*REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE*
*MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE*

*UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU*
*FACULTE DE GENIE ELECTRIQUE ET INFORMATIQUE*
*DEPARTEMENT INFORMATIQUE*

# *Mémoire de fin d'études*

***En vue d'obtention du diplôme de Master en informatique***

***Option : Conduite de Projets Informatiques***

# *Thème*

## *Traduction automatique de terminologie Informatique*

***Dirigé par :***
*M^me^ F.**AOUGHLIS***

***Réalisé par :***
- *M^elle^**OUZIA** Kahina*
- *M^elle^ **BOUGDOUR** Lamia*

***Membres du jury :***

| | | | |
|---|---|---|---|
| ***M^r^ R.Ahmed Ouamer*** | *Professeur* | *UMMTO* | *Président* |
| ***M^r^ S.Khemliche*** | *MAA* | *UMMTO* | *Examinateur* |
| ***M^r^ M S.Habet*** | *MAA* | *UMMTO* | *Examinateur* |

***Promotion 2014/2015***

# *Remerciements*

*La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui on voudrait témoigner toute notre reconnaissance.*

*Nous adressons d'abord notre gratitude à notre chère promotrice ; Madame AOUGHLIS Farida, pour sa patience, sa disponibilité et surtout ses judicieux conseils qui ont contribué à alimenter notre réflexion.*

*Nos plus sincères remerciements vont à tous les membres de jury qui nous ont fait l'honneur de juger notre travail.*

# *Dédicaces*

*Je dédie ce modeste travail à:*

*Mon très cher père qui a toujours su m'inculquer le sens de l'optimisme et de la confiance en soi.*

*Ma chère mère, ma source de tendresse, d'affection et de courage.*

*Ma sœur IMENE, qui m'a toujours guidée et consolée.*

*En particulier mon adorable frère KHALED.*

*Ma famille et mes proches.*

*Mes ami(e)s.*

*Ma très chère amie et binôme Kahina*

*Et à toute personne m'ayant fait part de son savoir.*

*Bougdour Lamia*

*Je dédie ce modeste travail à:*

*Mes très chers parents qui m'ont soutenu et encourager durant tout mon cursus*

*Mes très cher(e)s sœurs et frères ainsi que leurs époux et épouses. En particulier, mon grand frère "Kaci" merci pour ton soutien moral et ton assistance durant tout le long de ma formation! Que dieu te bénisse chère frère.*

*Mes adorables nièces et neveux!*

*Mon binôme et amie Lamia, je te dédie ce travail aussi en tout souhaitant un avenir radieux et plein de bonnes promesses!*

*Et enfin à tous ceux qui m'ont soutenu de près et de loin.*

*Ouzia Kahina*

# Table des matières

**Chapitre II: La traduction automatique**

**Chapitre III: Mots composés et terminologie informatique**

**Chapitre IV: Le système NooJ**

**Chapitre V : Ressources linguistiques et exemples**

# Liste des figures

# *Introduction*

# Introduction

La linguistique et l'informatique sont, à priori des disciplines indépendantes l'une de l'autre. En effet la première, cherche à comprendre le fonctionnement d'une langue et les processus du langage. La seconde vise à concevoir des machines portant le nom

« d'ordinateur » afin de traiter l'information.

Pourtant, elles unissent leurs efforts pour doter cette machine de réelles capacités de communication donnant naissance à une nouvelle discipline: « Le Traitement automatique des langues ». Cette dernière désigne toutes activités de développement de programmes et techniques informatiques, pour la manipulation automatique de données textuelles exprimées dans une langue donnée.

Les premières recherches du TAL ont porté sur la traduction automatique. Au cours des années 50, on entrevoyait la possibilité de fabriquer des logiciels adéquats qui permettraient à l'ordinateur de traduire un terme d'une langue par le terme équivalent dans une autre langue, la traduction automatique des débuts était donc essentiellement une traduction de mots. Par la suite, il devient possible de traduire des phrases et des textes d'une langue à une autre utilisant un ensemble de ressources linguistiques notamment les dictionnaires. En effet lors de la traduction, le moteur doit nécessairement recourir à la consultation des dictionnaires, comme dans le cas du système Nooj.

Actuellement, un grand intérêt est accordé aux langues de spécialités et à la construction de ressources linguistiques. Néanmoins dans Nooj, les dictionnaires spécialisés sont pratiquement inexistants. Nous nous intéresserons dans notre cas à une langue technique, celle de l'informatique. L'objectif de notre recherche, est de pouvoir développer une traduction automatique de terminologie informatique, allant de la langue française à la langue anglaise et vice-versa.

Afin de réaliser ce travail, nous avons jugé bon de répartir notre mémoire en cinq chapitres comme suit :

- Le premier chapitre porte sur le traitement automatique des langues dans lequel on a retracé son historique et son évolution, ainsi que ses domaines d'applications.
- Le deuxième chapitre, la définition de la traduction automatique, son principe de fonctionnement et ses difficultés sont abordés.
- Le troisième chapitre, présente, un ensemble de définitions : mot, terme, mot composé, synapsie, etc. Et une partie concernant la terminologie informatique.
- Le quatrième chapitre, décrit l'environnement de notre travail, NooJ. Ses différentes fonctionnalités tout en l'illustrant avec des exemples d'utilisation.

- Le cinquième  et dernier chapitre, sera consacré à la mise en place de toutes les ressources linguistiques, dictionnaires, grammaires syntaxiques, grammaires de flexions en vue de la traduction de texte de termes composés d’informatiques. Une démonstration du travail réalisé est présentée à l'aide d'exemples.

# *Chapitre I*

# *Le traitement automatique des langues*

# Chapitre I: Le traitement automatique des langues

## I.1 Introduction

Le traitement automatique des langues (T.A.L) est un domaine de recherches pluridisciplinaires, qui fait collaborer linguistes, informaticiens, logiciens, psychologues, documentalistes, lexicographes ou traducteurs, et qui appartient au domaine de l'Intelligence Artificielle (I.A.), Ceci fera l'objet de ce premier chapitre où on présentera l'historique de cette approche, et notamment les différents niveaux de traitement automatique des langues, son utilité ainsi que les difficultés.

## I.2 Le traitement automatique des langues : Une première définition [1]

Le traitement automatique des langues naturelles (TALN) a pour but la création de programmes informatiques permettant de traiter automatiquement les langues naturelles.
La langue naturelle désigne la langue parlée ou écrite par les être humains, par opposition aux langages artificiels, informatiques, mathématiques ou logiques, par exemple.
Le traitement porte sur les données linguistiques, les textes, codés dans une langue particulière. Sous cette dénomination générique, nous regroupons aussi les dialogues, écrits ou oraux, et des unités plus petites, comme les paragraphes ou les phrases.

D'une première approximation, le traitement est la transformation d'un objet d'entrée en un objet de sortie. Quand il porte sur la langue, le traitement peut être de deux types:

- Il peut agir sur des données linguistiques (des textes) pour les corriger, les condenser ou les traduire. Souvent, cette transformation comprend une étape intermédiaire qui vise à extraire des textes leur représentation, comme en (1): elle est appelée analyse des langues naturelles.

(1) Texte $\longrightarrow$ Représentation (texte1) $\longrightarrow$ Texte 2
Analyse Génération

Le terme « représentation » désigne toute traduction du texte dans un système autre que la langue naturelle et qui rend explicites des informations implicites dans le texte: un ensemble de mots clés, un arbre syntaxique, une formule logique, etc. Dans ce premier type de traitement, l'entrée est donc un texte et la sortie un nouveau texte ou une représentation de texte.

- Le TAL peut aussi faire l'opération inverse : il prend alors en entrée la représentation de texte, pour produire un texte en langue naturelle. Cependant, on ne dispose pas généralement de la représentation de textes mais de données brutes, comme des tableaux ou des tables, qu'il faudra d'abord traduire en une représentation de texte, comme en (2), cette opération est appelée génération des langues naturelles.

(2) Données brutes ⟶ Représentation (texte) ⟶ texte

Génération

Ce traitement automatique nécessite évidemment des outils divers que l'on peut grouper en trois catégories distinctes :

- **1ère catégorie:** Linguistique, ils décrivent les diverses connaissances relatives à la langue;
- **2ème catégorie**: Formels, ils expriment ces connaissances dans un formalisme qui convient à un traitement automatique ;
- **3ème catégorie:** Enfin, en informatique, ils utilisent cette description formelle des connaissances dans une application informatique concrète. Il ne faut donc pas s'étonner de la diversité du TAL, qui fait intervenir des recherches dans différents domaines:
  1- La linguistique informatique : qui développe des programmes de TAL, et définit, dans ce but, de véritables langages informatiques, spécialisés pour les applications du TAL;
  2- La linguistique : qui fournit des théories explicites du savoir linguistique;
  3- L'informatique : qui permet d'optimiser les algorithmes et les programmes de traitement, mais aussi de développer des techniques formelles (de démonstration ou de résolution de problèmes par exemple);
  4- Les mathématiques : qui étudient les propriétés formelles des outils de traitement et de théories ;
  5- L'intelligence artificielle (IA) : qui s'occupe de la représentation des connaissances et de leur utilisation.

## I.3 Pourquoi traiter automatiquement les langues naturelles? [1]

Deux raisons principales, justifient le TAL :

Tout d'abord sur un plan théorique, le TAL permet de vérifier les théories linguistiques, de mieux comprendre comment les humains communiquent entre eux. Dans ce but, il utilise l'ordinateur pour simuler les capacités humaines de compréhension et de production de la langue naturelle. Les résultats obtenus peuvent ensuite être comparés aux performances humaines et les théories sur lesquelles se fondent les simulations vérifiées.

Ensuite, sur un plan pratique, le TAL rend possible la construction de systèmes opérationnels qui débouchent sur des produits commerciaux. Dans le paragraphe suivant on étudiera les applications les plus importantes.

### I.3.1 La traduction automatique (TA)

Représente la première application du traitement automatique de langue. La traduction automatique se définit comme l'application de l'informatique à la traduction de textes oraux ou écrits d'une langue naturelle de départ (ou langue source) dans une langue d'arrivée (ou langue cible). Elle porte sur plusieurs intermédiaires: à une extrême, la traduction automatique assistée par l'homme(TAAH) (qualifiée aussi de traduction assistée par ordinateur (TAO) ou de traduction automatique assistée(TAA)), qui nécessite une intervention humaine; à l'autre, la traduction entièrement automatique.

Le terme de TA n'inclut donc pas les outils informatiques d'aide à la traduction ou aides informatisées à la traduction qui s'inscrivent dans le cadre de la traduction humaine assistée par ordinateur(THAO) et qui permettent au traducteur l'accès à des dictionnaires ou à des bases de données terminologiques, l'édition de texte, la gestion de glossaires, la constitution de concordances, etc. Ces concepts sont repris sous la dénomination générique de « traductique » qui recouvre l'ensemble des disciplines relatives à l'informatisation de la traduction.

### I.3.2 L'indexation automatique et la recherche documentaire

Une recherche documentaire automatique permet de retrouver automatiquement les informations, les documents pertinents, ou alors les références des documents, qui répondent à la requête d'un utilisateur. Comme la plupart des informations ont la forme de texte en langue naturelle, l'indexation automatique transforme ce dernier en une représentation pour le rendre accessible par la suite, lors de la recherche documentaire.

### I.3.3 Les systèmes multimédia

Ceux-ci sont considérablement développés ces dernières années et permettent non seulement le stockage d'informations sous diverses formes (textuelle, graphiques, audio, vidéo, etc.), mais aussi leur accès et leur manipulation à distance.

### I.3.4 Les systèmes de dialogue Homme-Machine

Au début des années soixante-dix, les interfaces en langue naturelle, qui permettent à l'être humain de communiquer avec les ordinateurs, ont connu un succès important. En effet, elles représentent une manière plus conviviale pour dialoguer avec une machine. Interrogation de banques de données en langue naturelle, commande de robots ou de machines ou encore dialogue avec des systèmes experts restent toujours les applications de prédilection du TAL.

### I.3.5 Les correcteurs

Parmi les systèmes moins ambitieux que ceux de TA : les correcteurs d'orthographe, de syntaxe et de style, qui se limitent à l'analyse du langage et aident l'utilisateur à corriger un texte. Outils très répandus, ils font déjà partie intégrante de la plupart des logiciels de traitement de texte.

### I.3.6 La génération de textes en langue naturelle

Cette application est utilisée afin de produire des textes en langue naturelle à partir de données non linguistiques, comme des graphiques, des schémas, des dessins, des données numériques, etc. Idéale dans des domaines tels la météorologie et la bourse où l'on dispose de données brutes, non linguistiques, qui doivent être traduites régulièrement dans des langues naturelles.

Malgré leur diversité, ces applications de TAL sont encore mal connues dans le monde francophone, en dépit d'efforts récents comme :

1- **La création d'observatoires francophones des industries de la langue :** Regroupés au sein du réseau international francophone des industries de la langue(RIOFIL), et de réseaux spécialisés dans le domaine, tels ceux de l'AUPELF-UREF, l'Agence francophone pour l'enseignement supérieur et la recherche (notamment le réseau lexicologie, terminologie et traduction (LTT) et le réseau francophone de l'ingénierie de la langue 'FRANCIL).
2- **L'organisation de colloques**, etc.

## I.4 Brève histoire du traitement automatique du langage naturel [1]

L'histoire du TAL se mêle à celle de la TA, qui représente l'une des premières applications informatiques non numériques. Depuis la fin de la Seconde Guerre mondiale, elle a connu un passé tumultueux qui a été souvent bien décrit et dont nous ne reprendrons ici que les faits essentiels.

### I.4.1 Années 50-66

Après la Seconde Guerre mondiale, le développement des ordinateurs favorise de nouveaux projets .Warren Weaver de la fondation Rockefeller, éminent mathématicien américain, influencé par les théories de Shannon, persuade des personnalités influentes que la traduction automatique est, pour les ordinateurs, une tâche aisée qui devrait nécessiter d'autres techniques que celles développées avec succès pour le décryptage des messages codés.

Il réussit ainsi à réunir des fonds considérables. Un intérêt stratégique énorme est en jeu : dans le contexte d'après-guerre, la possibilité de traduire automatiquement des textes

intéresse au plus haut point l'armée et les services d'espionnage. Les recherches commencent dans l'euphorie : les langues traitées se limitent au russe et à l'anglais. Cet enthousiasme, portant, décline rapidement : dans un compte-rendu de 1960 sur les progrès de la TA, Yehoshua Bar-Hillel, une des personnalités les plus convaincues des débuts, montre déjà qu'il est impossible de concevoir une traduction entièrement automatique de bonne qualité sans prendre en considération le sens, ce qui semble impossible à l'époque. Ce désaveu n'est pas étonnant et l'échec des recherches en TA qui s'en suivit trouve de nombreuses explications : tout d'abord, l'informatique en est à ses premiers balbutiements (les mémoires sont limitées et les programmes ne peuvent pas être distingués des données linguistiques) ; de plus, les méthodes de programmation sont primitives. Ensuite la linguistique est peu développée : le structuralisme qui fonde son étude sur des énoncés réalisés et ne s'intéresse pas aux problèmes de représentation du sens n'est pas directement utile pour les systèmes généraux de traduction automatique de l'époque (De Roeck, 1978, par exemple). Enfin, les chercheurs ne se rendent pas compte des problèmes que pose la traduction.

### I.4.2 Année 1966: le rapport ALPAC

En 1966, les autorités qui ont investi des sommes considérables dans les recherches en TA désirent des résultats concrets et chargent le comité ALPAC (Automatic Language Advisory Committee) d'évaluer les progrès de cette discipline. Le rapport, influencé par les nombreux critiques et sceptiques qui tournent en ridicule le domaine, est désastreux ; il conclut que les recherches dans leur état actuel ne sont pas rentables pour l'Etat Américain. Les subsides pour la TA sont donc coupés aux USA du jour au lendemain. Faute de moyens, les chercheurs se tournent vers d'autres directions qui permettent de diversifier les recherches du TAL:

**- Des recherches théoriques en syntaxe :** à la fin des années 60, les théories de Noam Chomsky sur les grammaires formelles et la grammaire générative transformationnelle ont un double impact. Tout d'abord, Noam Chomsky montre que la structure des phrases est assez régulière pour être décrite de manière formelle dans une grammaire qui rende compte de la compétence linguistique des être humains. En second lieu, il développe une théorie qui débouche sur un traitement purement syntaxique des langues : elle présuppose en effet que la syntaxe est indispensable pour extraire le sens de la phrase et que l'analyse syntaxique peut se faire indépendamment de toute analyse du sens.

Ces recherches soulèvent un nombre de réactions dans le milieu du TAL: la grammaire transformationnelle est-elle adéquate pour le TAL ? Est-il vraiment souhaitable de séparer les analyses syntaxiques et sémantiques et de faire l'analyse syntaxique d'une phrase, sans prendre en considération son sens, les connaissances du monde ou le contexte d'utilisation ? Certaines critiques de la théorie Chomskienne sont même plus profondes : on se demandait même si l'analyse syntaxique est

vraiment indispensable pour extraire le sens des phrases. Dans ce contexte, différentes études se tournent vers le problème de la représentation du sens, parallèlement à une recherche plus générale consacrée à la représentation des connaissances en Intelligence Artificielle (IA). Ces recherches aboutissent à des systèmes concrets dans le domaine qui vont témoigner de la possibilité de traiter automatiquement les langues naturelles, sans mettre nécessairement en œuvre la grammaire Générative transformationnelle.

Différents programmes sont en effet particulièrement représentatifs de l'époque post-Alpac : des systèmes de dialogue homme-machine, d'abord comme le programme ELIZA (Allen, 1987 ; Sabah 1988 et Weizebaum 1966 et 1967), où la machine joue le rôle d'un thérapeute et SHRDLU( Winograd, 1972), où l'humain dialogue une machine à propos d'un monde clos constitué de formes géométriques (des pyramides, des cubes, etc.) qu'il peut manipuler ; des systèmes d'interrogation de banques de données, ensuit, comme par exemple, le système LUNAR de woods (Woods,1968).

Reposant sur des bases théoriques très différentes, ces premiers systèmes de TAL permettent un certain nombre de conclusions qui paraissent contradictoires (Ritchie et Pullman, 1991) :

- Il est possible de traiter les langues sans mettre en jeu des connaissances linguistiques. Par exemple Eliza dialogue sans comprendre. Ce système n'utilise pas de connaissances linguistiques, mais une banque de données qui contient un stock de phrases indexées par des mots-clés dont le poids est le plus élevé et de générer, en guise de réponse, la phrase qui lui correspond ;

- On peut aussi obtenir de bons résultats, en mettant en œuvre la grammaire transformationnelle de Chomsky.

- Enfin, on peut obtenir des résultats similaires sans mettre en œuvre la théorie de la grammaire transformationnelle.

En fait, il n'y a pas de contradiction, les systèmes des années soixante-dix ne sont que des mises en œuvre partielles de théories, qui ne prouvent pas leur validité. De plus, ils témoignent bien de la difficulté d'évaluer la qualité d'un programme à son résultat ou son intelligence à son comportement : ce n'est pas parce qu'un programme fonctionne que la théorie qui le sous-tend est adéquate.

Quoi qu'il en soit, ces programmes, en montrant l'intérêt du TAL, vont contribuer considérablement à son développement.

### I.4.3 Année 1975 et suivantes

A partir des années soixante-quinze, le TAL va en effet prendre de plus en plus d'importance sous l'effet de différents facteurs :

1- le regain d'intérêt pour le TAL, grâce au succès de programmes comme ELIZA, SHRDLU et LUNAR (qui font oublier l'échec de la traduction automatique) ;
2- Un effort académique et politique : action de certaines sociétés, comme l'association Of Computational Linguistics (ACL) ou les observations des industries de la langue, création de programmes d'études dans plusieurs universités (voir, pa exemple, les Survey of Computational Courses, édités par l'ACL en 1986 et 1994 ainsi que la revue T.A.L, vol 37, 1996, dédiée à l'enseignement du TAL), organisation de conférences périodiques et d'écoles d'été, etc ;
3- Le développement des techniques informatiques (logiciel et matériel) et celui de nouvelles théories linguistiques, en réaction à celle de l'école Chomskienne qui convient mieux au TAL ;
4- Le développement de l'informatique qui envahit tous les secteurs de la société ;
5- La nécessité de traiter et de gérer une masse toujours croissante d'informations multilingues;

A cet égard, 1975 est une date importante pour le développement du TAL en Europe. La communauté Européenne, qui doit faire face à l'accroissement alarmant du nombre de traductions, entrevoit la possibilité de recourir à la traduction automatique. L'année suivante, elle déclenche un plan d'action dont le but est de coordonner différents projets qui traitent du multilinguisme et, notamment, de TA, quand elle annonce en 1976 l'installation d'un système de TA commercial, nommé Systran (voir Hutchins et Somers, 1992)[1], la traduction automatique se fait connaitre du grand public et suscite à nouveau l'intérêt des firmes privées : elle est désormais une application rentable et les systèmes commerciaux se multiplient.

Depuis, différents programmes de TAL se succèdent, contribuant ainsi au développement des industries de la langue. Citons, à titre d'exemple, les programmes européens EUROTRA (1986-1992) de traduction automatique entre les neuf langues de la communauté européenne. Linguistic Research & Engineering LRE (1994/1998) pour le développement d'applications du TAL et de méthodes, outils et ressources de base et enfin, Language Engineering LE (1994-1998) qui continue les activités développées dans le cadre du programme LRE.

[1] Revue intitulée « An Introduction to Machine Translation » de W. John Hutchins and Harold L. Somers. London, Academic Press,1992.

## I.5 Les niveaux de traitement automatique des langues [2]

Selon le découpage méthodologique classique dans le domaine et en linguistique, les grands domaines du TAL sont :

- **Le lexique :** qui concerne l'étude de la formation des mots et de leurs variations de forme ;
- **La syntaxe:** qui s'intéresse à l'agencement des mots et à leurs relations structurelles dans un énoncé ;
- **La sémantique:** qui se consacre au sens des énoncés ;
- **La pragmatique:** qui prend en compte le contexte d'énonciation.

Ainsi on présente ces domaines sous l'angle de l'analyse automatique des textes, commençant par le lexique, la syntaxe, la sémantique et enfin la pragmatique.

### I.5.1 Le lexique

D'un point de vue informatique, un texte représente une chaîne de caractères. Son analyse commence par une première étape : la reconnaissance, d'unités linguistiques de base, les mots, et la mobilisation des informations associées.

Initialement le lexique, est la liste des mots de la langue, et qui associe à chaque mot les informations linguistiques correspondantes : catégorie syntaxique, traits morphosyntaxiques (genre, nombre, etc.), etc. Plusieurs phénomènes amènent à préciser cette définition du lexique :

**1-** Un mot peut exister sous plusieurs formes : en français, formes fléchies des noms, adjectifs, etc., conjugaison des verbes. On peut alors considérer une *forme canonique*, ou lemme, pour chaque mot, qui sert d'entrée dans le lexique pour l'ensemble de ses formes fléchies (singulier pour le nom, masculin singulier pour l'adjectif, infinitif pour le verbe).

**2-** Un mot peut avoir plusieurs sens (*polysème*), comme par exemple *« avocat », « coup », « livre »* exemples ; selon le cas, plusieurs entrées ou sous entrées sont alors distinguées.

**3-** Plusieurs mots peuvent se trouver partager une forme commune (*homographes*):

- *« montre »* est une forme du nom *« montre »* aussi bien que du verbe *« montrer »* ;
- *« pu »* est le participe passé du verbe *« pouvoir »* mais aussi de *« paître »*.

**4-** Un mot peut être construit à partir d'un autre : par dérivation (*« penser» -> « pensable » -> « impensable »*) ou par composition (*« compter» + « gouttes» -> « compte-gouttes », « un » + « jambe » -> « Unijambiste », « sclérose » + « artère » -> « artériosclérose »*).

**5-** Enfin, pour de multiples raisons, tous les mots possibles d'une langue ne peuvent être répertoriés dans un lexique. D'une part, les noms propres constituent un inventaire ouvert. D'autre part, de nouveaux mots sont créés régulièrement (néologie) soit par dérivation et composition, ou encore par siglaison, abréviation, emprunt, etc.

Pour commencer l'analyse, la chaîne de caractères d'entrée doit utiliser un encodage déterminé (typiquement, pour le français, l'encodage ISO-latin-1), les caractères de contrôle (fin de ligne, etc.) étant eux aussi normalisés. On élimine généralement les caractères non répertoriés.
Par la suite on procède à la segmentation de la chaîne d'entrée en unités élémentaires (en anglais, *tokens*). Différents choix peuvent être effectués à cette étape, selon les séparateurs choisis : tous les caractères non alphabétiques (espaces, apostrophes, tirets...) ou les espaces seulement ; et selon que l'on prend en considération les « mots composés » (*« pomme de terre »* = une unité) ou pas. En tout état de cause, on est généralement amené à distinguer la notion d'unité minimale (« token ») et celle de mot (associé à une information lexicale).

## I.5.2 La syntaxe

Comment repérer quels mots fonctionnent ensembles dans une phrase ? Un premier niveau de modélisation consiste à constituer des classes de mots (catégories syntaxiques, parties du discours) possédant un fonctionnement similaire: Nom (N), Verbe (V), Adjectif (A), etc.
Certaines unités, par accident (homographes: *« la »*, *« est »*) ou de façon plus systématique (*« normale »*: A -> N, *« coronarographie »* N -> *« coronarographier »* V -> *« coronarographie »* V), peuvent être *ambiguës* entre plusieurs catégories (ambiguïté catégorielle ou lexicale).

Les relations syntaxiques entre les mots d'une phrase peuvent se représenter de plusieurs façons. Le modèle en constituants considère des groupes de mots, ou syntagmes, généralement centrés sur un mot de tête (N, V, etc.), et les modélise par des catégories spécifiques (syntagme nominal ou SN, syntagme verbal ou SV, syntagme adjectival ou SA, etc.). Ces syntagmes peuvent eux-mêmes être éléments d'autres syntagmes, et la structure d'une phrase est alors un *arbre de constituants* **(Figure I.5.2(a)).**

Le modèle en dépendance considère directement les mots de tête (recteurs, ou régissant), et leur attache les mots qui en dépendent (régis).

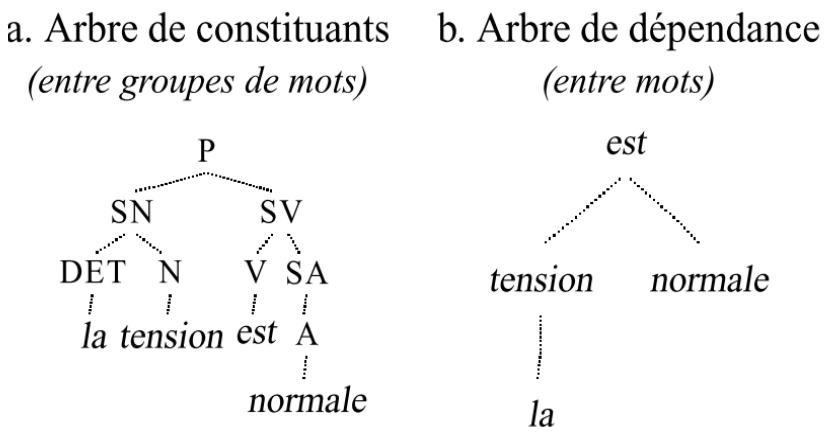


**Figure I.5.2 (a) Représentation syntaxiques d'une phrase [2]**

La structure d'une phrase est alors un arbre de dépendance. Des équivalences existent entre les deux modèles.

Même sans ambiguïté lexicale, une phrase peut donner lieu à plusieurs structures syntaxiques (ambiguïté structurelle).
Un exemple classique est la phrase *« je vois un homme avec un télescope »* (**Figure I.5.2(b)**) dans laquelle *« avec un télescope »* peut désigner la manière dont je vois l'homme (2a, attachement au verbe *« vois »*, complément circonstanciel de manière) ou au contraire une caractéristique de l'homme (2b, avec un attachement au nom *« homme »*, complément de nom). Des informations sémantiques, voire pragmatiques comme ce serait le cas dans cet exemple, sont nécessaires pour déterminer l'interprétation la plus appropriée.

Des relations plus précises entre mots ou syntagmes sont utiles à l'interprétation des phrases. Les relations grammaticales classiques (sujet-verbe, verbe-objet, verbe-objet-indirect, nom-modifieur, etc.) permettent de représenter la fonction des groupes de mots les uns par rapport aux autres. Les relations entre pronom et antécédent, et plus généralement entre anaphore (pronom, mais aussi nom) et antécédent, mobilisant encore davantage sémantique et pragmatique, sont très utiles en recherche d'information.

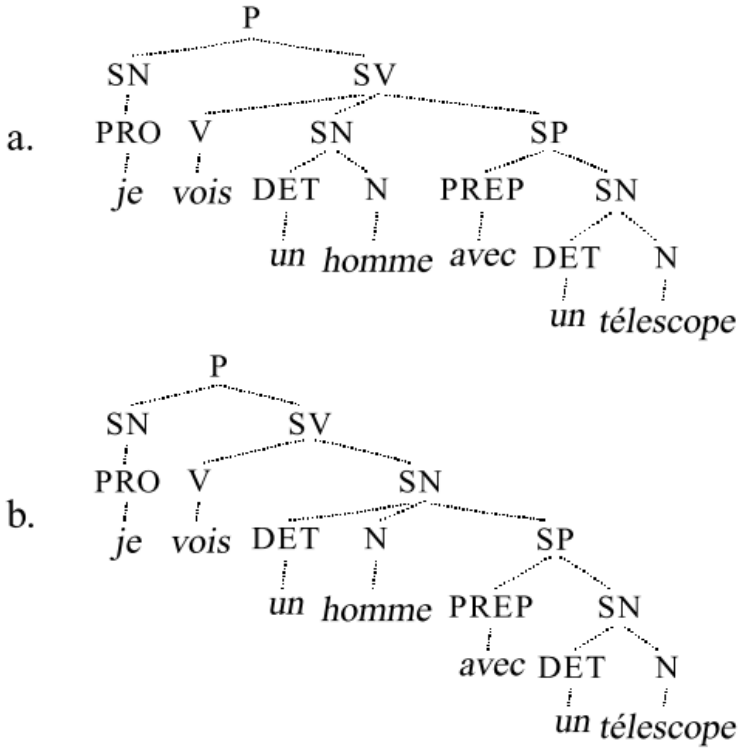


**Figure I.5.2 (b) Ambigüité structurelle**

Les propriétés intrinsèques des mots restreignent le type de relations syntaxiques qu'ils peuvent avoir. C'est en particulier le cas des verbes, qui régissent ou sous-catégorisent de zéro à trois ou quatre « arguments » :
*« Il pleut. »  pleuvoir()*
*« John marche. »  marcher(X)*
*« John mange une pomme. »  manger (X, Y)*
*« John offre un cadeau à Marie. »  offrir (X, Y, Z)*
*« John interdit à Jacque de sortir. »  interdire(X, Y, Z)*

### I.5.3 La sémantique

Pour la sémantique le premier niveau de modélisation est de même que pour la syntaxe, il consiste à constituer des classes de mots (catégories sémantiques). Ces classes regroupent des mots dont le sens est voisin, ou au minimum (pour des classes générales) des mots qui possèdent certaines propriétés sémantiques communes. Cependant, si en syntaxe on arrive à s'accorder sur des jeux de catégories relativement consensuels (il s'agit d'une vue d'ensemble ; de près, le tableau est beaucoup plus polychrome), en sémantique aucune classification universelle n'existe (la constitution d'une classification universelle risque même d'être théoriquement impossible). Les classifications qui pourront être utilisées reflètent nécessairement un point de vue, une prise de position culturelle ou ontologique spécifique.

### I.5.3.1 Mot à plusieurs sens

Un mot, même syntaxiquement non ambigu, pourra posséder plusieurs sens. Par exemple, on pourra distinguer l'*« artère »*–vaisseau sanguin de l'*« artère »*–avenue, même si le second est étymologiquement un sens figuré du premier. Le contexte permet en général de déterminer quel sens est à l'œuvre dans un énoncé.
Les mots d'une langue entretiennent un réseau riche de *relations sémantiques paradigmatiques* : hyperonymie / hyponymie (*« vaisseau »*/*« artère »*), méronymie (partie d'un tout : *« vaisseau »* / *« système cardiovasculaire »*), antonymie (*« malin »* / *« bénin »*) et autres contraires, etc.

### I.5.3.2 Relations grammaticales

Dans un énoncé, les relations grammaticales sont le support de *relations sémantiques syntagmatiques*. Par exemple, les différents actants d'un événement jouent différents *rôles thématiques* : agent, thème, source, destination, etc. Ainsi, dans *« Jean donne un livre à Marie. »*, les rôles par rapport à l'événement *« donne »* pourront être :
*« Jean/agent, source donne un livre/thème à Marie/destination. »*. Un mot qui désigne un événement possède des propriétés combinatoires restreintes : il sélectionne comme actants certains types de mots (*restrictions de sélection*). Ces types restreints peuvent être exprimés en termes de classes sémantiques. On pourra par exemple poser pour le verbe *« donner quelque chose à quelqu'un » donner (animé)*, ou encore pour *« interdire» interdire (animé, animé, événement)*.

### I.5.3.3 La représentation sémantique

La représentation sémantique finale que l'on vise à associer à un énoncé dans un système de TAL dépend de l'objectif de ce système. Cet objectif peut être l'extraction d'informations spécifiques, comme c'est le cas dans les tâches définies dans les campagnes MUC d'évaluation de systèmes d'analyse de textes. Par exemple, l'évolution des postes d'une personne dans une ou plusieurs entreprises (**Figure I.5.3.3(a)**) était l'une des tâches de la campagne MUC6. Un éventail d'informations plus large peut aussi être recherché. La représentation doit alors être plus complète, comme dans le système MENELAS. La (**Figure I.5.3.3(b))** montre une représentation de la phrase *« Patient âgé de 62 ans, hospitalisé pour angor spontané à répétition. »*.

<Template-93-1> :=
Doc_Nr: "93"
Content: <Succession_Event-93-1>

<Succession_Event-93-1> :=
Succession_Org: <Organization-93-1>
Post: "president"
In_And_Out: <In_And_Out-93-1>
Vacancy_Reason: OTH_UNK

<In_And_Out-93-1> :=
Io_Person: <Person-93-1>
New_Status: IN
On_The_Job: YES
Other_Org: <Organization-93-1>
Rel_Other_Org: SAME_ORG

<Organization-93-1> :=
Org_Name: "Prime Corp."
Org_Descriptor: "the financial-services company"
Org_Type: COMPANY

<Person-93-1> :=
Per_Name: "John Simon"

**Figure I.5.3.3 (a) : Représentation à la MUC[2]**

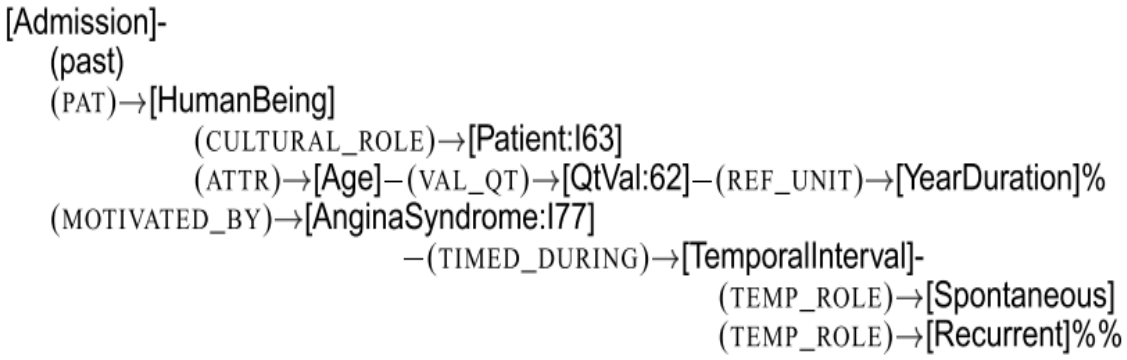


**Figure I.5.3.3(b) : Une représentation en Graphes Conceptuels [2]**

Les formalismes de représentation employés sont en général issus de l'Intelligence Artificielle, comme les logiques de description et les graphes conceptuels.

## I.5.4 La pragmatique

L'interprétation d'un énoncé dépend de son contexte. Dès que l'on veut traiter plus d'une phrase (et même pour une seule phrase), cette dimension intervient.

Le *co-texte* désigne le texte qui précède (et suit) la phrase courante. Deux facteurs concourent à faire qu'une phrase s'insère bien dans un texte :

1- **La cohésion** : régit la continuité du texte. Elle est assurée par l'emploi d'anaphores, l'homogénéité du thème, un emploi judicieux d'ellipses, etc.

2- **La cohérence :** détermine l'intelligibilité du texte. Elle s'appuie sur des structures de discours ainsi que sur les relations causales, temporelles, etc., entre les événements décrits.

Au-delà du texte lui-même, les conditions d'énonciation et les connaissances partagées complètent le contexte d'un énoncé. L'interprétation devra donc faire appel à des connaissances sur le monde (scénarios, plans, etc.). L'identification de structures de discours (structure de dialogue, structure argumentative, etc.) est également nécessaire selon le type de texte. De façon générale, une représentation de la situation décrite par un énoncé demande d'effectuer des inférences, ces dernières consistent en l'interprétation, pour un contenu propositionnel donné, une signification supérieure à la somme de ce qui a été simplement énoncé, en faisant intervenir des éléments de contexte intra et extratextuels issus à la fois de l'entourage linguistique et de l'univers de référence des interlocuteurs.

## I.6 Les outils du TAL [3][4]

On peut répartir les outils utilisés dans le traitement automatique des langues comme suit :

### I.6.1 Famille des correcteurs

Représente les logiciels permettant d'assurer les différentes fonctions de correction, elle se divise en trois parties :

**I.6.1.1 Les correcteurs grammaticaux :** permettent de corriger la structure grammaticale d'un texte.

**I.6.1.2 Les correcteurs orthographiques/lexicaux :** détectent les erreurs orthographiques d'un texte et proposent la correction.

**I.6.1.3 Les correcteurs stylistiques :** L'ensemble des programmes qui corrigent le style d'un texte. Ils éliminent les répétitions, constatent les mêmes ensembles de mots souvent présents dans le texte, signalent les phrases trop longues, etc.

On peut citer comme exemple d'outils le correcteur orthographique et grammatical "Cordial", conçu par la société toulousaine Synapsys Developpement, qui est le fruit de l'effort d'une équipe d'une dizaine de développeurs et linguistes. Le mot "Cordial" est un acronyme pour "CORrecteur D'Imprécisions et Analyseur Lexico-syntaxique".

### I.6.2 La famille des étiqueteurs

Constitue les programmes et les dispositifs ayant pour fonction d'associer des informations (étiquettes) distinctives à des mots, phrases, et passages d'un texte ou document (exemple l'étiqueteur « hybridtagger ») pour le français. Elle est composée de trois catégories :

**I.6.2.1 Les étiqueteurs morphosyntaxiques :** permettent d'ajouter ou d'associer des informations morphologiques et syntaxiques aux mots d'un corpus, d'un texte ou d'un document écrit ou orale.

**I.6.2.2 Les étiqueteurs prosodiques :** L'ensemble d'outils permettant d'ajouter ou d'associer des informations prosodiques aux mots d'un corpus, d'un texte ou d'un document. Désormais aux corpus oraux.

**I.6.2.3 Les étiqueteurs sémantiques :** Qui permettent d'ajouter ou d'associer des informations sémantiques aux mots d'un corpus, d'un texte ou d'un document écrit ou orale.

On cite comme exemple CONNEXOR, étiqueteur multilingue (Sept langues dont l'anglais et le français). Il est possible de choisir entre l'étiquetage FDG (Functional Dépendance Grammar), et l'étiquetage avec FDG Lite , plus léger (absence des catégories genre et nombre et des fonctions syntaxiques).

### I.6.3 La famille des outils de reconnaissance de données à partir de corpus

**I.6.3.1 Les outils de reconnaissance de caractères imprimés**: Procédé qui transforme les textes imprimés en code textuel, les rendant donc lisible par une machine.

## I.7 Les applications du TALN [5]

Concernant les applications, la demande de TALN provient de deux tendances :

1- D'une part la nécessité de concevoir des interfaces de plus en plus ergonomiques,
2- D'autre part la nécessité de pouvoir traiter (produire, lire, rechercher, classer, analyser, traduire) de manière de plus en plus « intelligente» les informations disponibles sous forme textuelle, de manière à pouvoir résister à leur prolifération exponentielle.

Les applications des techniques de TAL sont donc nombreuses et variées. Ces dernières ont été regroupées en trois grandes familles, qui correspondent aux aides à la lecture de documents, aux aides à la production de documents, et enfin aux interfaces homme-machines. Ces familles d'applications sont détaillées dans les paragraphes qui suivent.

### I.7.1 Le traitement documentaire

Les applications les plus immédiates du TALN sont celles qui visent à faciliter le traitement par l'humain des immenses ressources disponibles en langage naturel, comme par exemple :

**I.7.1.1 La traduction automatique** : Cette application, qui a historiquement suscité les premiers efforts de recherche en TALN, reste un enjeu économique et politique de première importance. Il est important de noter que même si la traduction complète indépendante du domaine est encore hors d'atteinte, on peut obtenir de bons traducteurs spécialisés (domaines techniques), qui constituent un moyen efficace de préparer l'intervention manuelle d'un traducteur. Il existe également des environnements de travail fournissant des ressources lexicales (dictionnaires bilingues étendus) pour l'aide à la traduction.

Aujourd'hui, de nombreux systèmes commerciaux sont disponibles (Systran, Logos, Metal, Aleth-Trad...), et certains moteurs de recherche proposent une traduction automatique des pages html.

**I.7.1.2 La recherche de documents :** La prolifération des outils de recherche documentaire sur la toile, qui traitent quotidiennement des millions de requêtes, montrent bien l'importance de la demande en la matière. Les performances de ces moteurs témoignent du chemin qu'il reste à parcourir dans ce domaine.
De plus en plus d'outils fournissent également des moyens de recherche spontanée d'adresses potentiellement intéressantes (à partir de profil utilisateurs), ou encore de surveillance automatique des publications dans des domaines donnés.

**I.7.1.3 Le routage :** Le classement ou l'indexation automatique de documents électroniques sont des variantes applicatives du paradigme de la recherche documentaire.
Plus complexe est la tâche de trouver (ou de produire à la demande) des réponses précises aux questions de l'utilisateur (tâche de "question-réponse").

**I.7.1.4 La lecture automatisée de documents :** par exemple pour les stocker dans des structures formelles de données, ou pour en extraire des résumés ;

**I.7.1.5 L'analyse d'un corpus de documents relatifs à un thème donné :** Comme (histoire, stylométrie, veille technologique,etc). Une application typique de ce domaine consiste à fournir des outils de visualisation et d'exploration dynamique de champs disciplinaires (scientifiques, par exemple).
Pour ce qui concerne les outils d'indexation et de recherche documentaire tels que ceux qui sont disponibles sur la toile, les techniques actuellement mises en œuvre sont essentiellement statistiques, et ne font appel, pour des raisons évidentes de coût de traitement, qu'aux outils de « bas-niveau » du TAL : segmentation et lemmatisation. La lemmatisation consiste à retrouver le lemme d'une forme fléchie, c'est-à-dire à lui retirer son ou ses suffixes flexionnels, c'est donc une forme très simplifiée d'analyse morphologique.

## I.7.2 La production de documents

Si autant de documents électroniques sont aujourd'hui disponibles, c'est bien que quelqu'un les a écrits. Dans le domaine de l'aide à la production de texte (la génération de textes), les applications du TALN sont également nombreuses :

**I.7.2.1 Les claviers** : « auto-correcteurs » (par exemple pour les handicapés) ;

**I.7.2.2 La reconnaissance optique de caractères** : De nombreux systèmes commerciaux sont aujourd'hui disponibles, avec des performances très satisfaisantes : Recognita, Omnipage, ScanWorX... ;

**I.7.2.3 Les correcteurs d'orthographe ou de syntaxe :** De tels correcteurs sont aujourd'hui disponibles dans la majorité des systèmes de traitement de texte commerciaux, avec des performances variables suivant les mécanismes de correction

mis en œuvre, qui vont de la recherche lexicale tolérante à l'analyse syntaxique partielle ou complète de la phrase.

**I.7.2.4 Les correcteurs « stylistiques » :** ou les aides intelligentes à la rédaction intégrant des thésaurus, des connaissances sur les « bonnes » pratiques rédactionnelles, etc.

**I.7.2.5 L'apprentissage :** assisté par ordinateur des langues naturelles.

**I.7.2.6 La génération automatique de documents :** à partir de spécifications formelles. En fait, de nombreux secteurs d'activité impliquent la production massive de textes très stéréotypés à partir de spécifications plus ou moins formelles (textes juridiques, compte-rendu d'exploration d'une base de données, rapports d'analyses statistiques, documentations techniques, etc).

On retrouve dans ces applications la même dialectique que dans les applications destinées à faciliter la gestion de documents. D'un côté des applications à large couverture, qui utilisent essentiellement des ressources lexicales, avec des fonctions d'accès tolérant (permettant la correction d'erreurs) au lexique : c'est le cas des applications qui tournent autour de la correction orthographique. De l'autre, des applications qui intègrent des mécanismes de traitement de plus haut niveau (typiquement la génération), mais qui ne fonctionnent que pour des domaines beaucoup plus restreints.

## I.7.3 Les interfaces naturelles

Dernier domaine d'application, qui est sans doute celui dans lequel la demande de traitement linguistique est la plus forte, le domaine des interfaces naturelles (i.e. en langage naturel) telles que :

**I.7.3.1 L'interrogation en langage naturel de bases de données :** (traduction langage naturel SQL) ou de moteurs de recherche sur la toile. De multiples applications de ce type commencent à se mettre en place sur la toile.

**I.7.3.2 Les interfaces vocales :** qui mettent en œuvre de manière variable suivant les applications des modules de reconnaissance de parole, synthèse de parole, génération et gestion de dialogue, accès aux bases de connaissance, chacun de ces modules demandant des traitements spécifiques (désambiguïsation morpho-syntaxique et identification de syntagmes pour la synthèse, grammaires stochastiques pour la reconnaissance de la parole.).

Les premiers systèmes « grand-public » de dictée vocale commencent à arriver sur le marché (Via Voice, le produit d'IBM, Dragon Dictate, et bien d'autres encore), et l'intégration dans Windows d'une API de traitement de la parole devrait, dans les années qui viennent, faire littéralement exploser le marché des technologies vocales. De nombreux services commerciaux existent déjà, ou sont proches de la commercialisation qui font appel à l'ensemble de ces techniques (reconnaissance-dialogue-synthèse), pour des applications diverses telles que les ordinateurs «main-libre », la lecture téléphonique de courriers électroniques, la réservation de billets (train, avion), la liste est pratiquement infinie.

## I.8 Avantages du traitement automatique de langue [6]

Dans le domaine des interfaces Homme-machine, on peut citer trois avantages qui militent en sa faveur:

1- **Peu d'apprentissage requis** : ce point est particulièrement important pour les services devant être utilisés par des personnes peu habituées aux outils informatiques. Il est beaucoup plus simple (naturel) pour une telle personne de rechercher un service en français courant, plutôt qu'en navigant à l'intérieur d'une arborescence.
2- **Concision :** généralement, la langue naturelle est plus concise pour interroger une base de données que tout formalisme informatique. On peut pour s'en convaincre comparer la formulation en français d'une requête avec sa transcription en SQL par exemple, qui fera souvent apparaître énormément de parenthèses et opérateurs booléens.
3- **Adapté aux domaines conceptuels complexes :** le traitement automatique des langues est plus adapté que l'interface classique en arborescence ou en interrogation de moteur de recherche pour les domaines dont on peut faire le tour conceptuellement, mais qui ne peuvent pas être réduits à une nomenclature simple et présentent une multitude de formulations potentielles. C'est par exemple le cas de l'annuaire des professionnels. Une même activité peut-être désignée de multiples façons et être dans plusieurs professions, ce qui rend impossible la représentation de l'ensemble dans une arborescence figée. Aussi, le système doit être capable d'interpréter la formulation du métier telle proposée par l'utilisateur.

## I.9 Les difficultés du TALN : ambigüe et l'implicite [5]

Les difficultés que l'on rencontre en TALN sont principalement de deux ordres, et ressortent soit de l'ambiguïté du langage, soit de la quantité d'implicite contenue dans les communications naturelles, on détaillera dans ce qui suit les deux points évoqués :

### I.9.1 Ambigu

Le langage naturel est ambigu, et ce à quelque niveau qu'on l'appréhende. Cette ambiguïté, loin d'être marginale, est un de ses traits caractéristiques. On peut d'ailleurs voir là le résultat d'un compromis inévitable entre, d'un côté une capacité d'expression quasi illimitée, et de l'autre des contraintes liées à la limitation des ressources physiologiques mises en œuvre (taille de la mémoire à long et court-terme, densité de l'espace phonétique, contraintes articulatoires, etc.). Cette ambiguïté se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement, comme en témoignent les exemples suivants :

1- Ambiguïté des graphèmes (lettres) dans le processus d'encodage orthographique : comparez la prononciation du i dans lit, poire, maison ;

2- Ambiguïté des terminaisons dans les processus de conjugaison et d'inflexion : ainsi un /s/ final marque à la fois le pluriel des noms, des adjectifs, et la deuxième (parfois également la première) personne du singulier des formes verbales ;

3- Ambiguïté dans les propriétés grammaticales et sémantiques (i.e. associées à son sens) d'une forme graphique donnée : ainsi manges est ambigu à la fois morpho-syntaxiquement, puisqu'il correspond aux formes, indicative et subjonctive du verbe manger), mais aussi sémantiquement. En effet, cette forme peut aussi bien référer (dans un style familier) à un ensemble d'actions conventionnelles (comme de s'assoir à une table, mettre une serviette, utiliser divers ustensiles, ceci éventuellement en maintenant une interaction avec un autre humain) avec pour visée finale d'ingérer de la nourriture (auquel cas il ne requière pas de complément d'objet direct) ; et à l'action consistant à effectivement ingérer un type particulier de nourriture (auquel cas il requiert un complément d'objet direct), etc. Comparez en effet :

(1) (a) Demain, Marie mange avec ma sœur.
(b) Marie mange son pain au chocolat.

4- Ainsi que les déductions que l'on peut faire à partir de ces deux énoncés : de (a), on peut raisonnablement conclure que Marie sera assise à une table, disposera de couverts, tout ceci n'est pas nécessairement vrai dans le cas de l'énoncé (b).

5- Ambiguïté de la fonction grammaticale des groupes de mots, illustrée par la phrase :

(2) il poursuit la jeune fille à vélo.

6- Dans cet exemple à vélo est soit un complément de manière de poursuivre (et c'est il qui pédale), soit un complément de nom de fille (et c'est elle qui mouline) ;

7- Ambiguïté de la portée des quantificateurs, des conjonctions, des prépositions. Ainsi, dans:

(3) Tous mes amis ont pris un verre

8- On peut supposer que chacun avait un verre différent, mais dans:

(4)Tous les témoins ont entendu un cri

9- Il est probable que c'était le même cri pour tous les témoins.

10- Ambiguïté sur l'interprétation à donner en contexte à un énoncé. Comparez ainsi la « signification » de non, dans les deux échanges suivants :

(5) (a) Si je vais en cours demain ? Non (négation)
(b)Tu vas en cours demain ! Non ! (j'y crois pas)

### I.9.2 Implicite

L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, sensément dotés d'une connaissance du monde et de son fonctionnement telle que l'immense majorité des éléments de contexte nécessaires à la désambiguïsation mais aussi à la compréhension d'un énoncé naturel peuvent rester implicites. La situation change du tout au tout dès qu'une machine tente de s'insérer dans un processus de communication naturel avec un humain : la machine ne dispose pas de cette connaissance d'arrière-plan, ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible, si l'on ne dispose pas de bases de connaissances additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique).

Un exemple typique est la désambiguïsation du référent du pronom personnel 'il' dans les trois énoncés suivants :

Le professeur envoya l'élève chez le censeur parce qu' ... :

(6) a. ... il lançait des boulettes (il réfère probablement à l'élève)
b. ... il voulait avoir la paix (il réfère probablement au professeur)
c. ... il voulait le voir (il réfère probablement au censeur)

En l'absence de telles connaissances, bien d'autres problèmes de compréhension deviennent pratiquement insurmontables : pensez par exemple aux ellipses, aux métaphores, et, plus généralement, aux figures de style.

Fort heureusement, il existe de nombreuses applications pour lesquelles ces difficultés peuvent être, dans une large mesure, circonscrites. Dès lors, en effet, que l'on restreint le cadre des textes analysés à un sous domaine particulier (textes juridiques, textes scientifiques, serveur d'information spécialisé dans les informations sportives...), il devient possible d'une part d'ignorer un grand nombre d'ambiguïtés, en particulier sémantiques (par exemple dans le contexte de textes juridiques, on pourra probablement négliger la possibilité qu'un avocat marron désigne un fruit un peu trop mûr) ; et d'autre part de représenter formellement un grand nombre des connaissances nécessaires à la compréhension des énoncés du domaine considéré. En fait, certains domaines d'activité ou contextes d'interactions spécifiques semblent restreindre de manière drastique l'ensemble des énoncés possibles (ou acceptables), simplifiant de manière considérable le traitement de ces véritables sous-langages par une machine.

## I.10 Conclusion

L’étude du langage naturel et des mécanismes nécessaires à la mise en œuvre à son traitement automatique par des machines est un domaine d’études foisonnant, et riche en applications potentielles ou émergentes.

De nombreux progrès restent à accomplir pour mieux comprendre cette faculté et pour bâtir des systèmes capables de soutenir la comparaison avec l’humain, mais l’état des connaissances permet aujourd’hui de proposer de nombreuses solutions efficaces à des problèmes et des demandes réels.

# *Chapitre II*

## *La traduction automatique*

# Chapitre II: La traduction automatique

## II.1 Introduction

La conception et la mise au point de l'ordinateur au cours des années 50 a suscité l'enthousiasme et permis de grands espoirs. On entrevoyait la possibilité de fabriquer des machines capables de résoudre les problèmes de la traduction. Il suffirait simplement de concevoir les logiciels adéquats. Ceux-ci permettraient à la machine de traduire un terme d'une langue par le terme équivalent dans l'autre langue, mais avant d'entamer un chapitre portant sur la traduction, il convient de définir d'abord ce terme.

## II.2 Définition générale de la traduction

- Selon **Larousse**[2] :
  TRADUCTION : action de traduire, de transposer dans une autre langue.
  TRADUIRE : faire passer un texte d'une langue dans une autre.

- Selon **Robert**[3] :
  TRADUCTION : texte ou ouvrage donnant dans une autre langue l'équivalent du texte original qu'on a traduit.
  TRADUIRE : faire ce qui était énoncé dans une langue le soit dans une autre, en tendant à l'équivalence sémantique et expressive des deux énoncés.

Cette définition est très simple, mais on constate que celle prise du dictionnaire Robert est néanmoins plus précise et évoque déjà l'équivalence sémantique et expressive. Cependant la définition classique du terme « traduction » peut être donnée comme suit :

**II.2.1 définition classique :** la définition classique acceptée par chaque traducteur est la suivante : « la traduction est l'opération qui consiste à transposer les informations contenues dans un texte source vers un texte cible ».

## II.3 Généralités sur la traduction automatique [7]

L'automatique peut être défini comme étant l'ensemble des disciplines scientifiques et des techniques utilisées pour l'automatisation d'un processus, autrement dit, la conception et l'emploi des systèmes automatiques.
La traduction automatique ou TA est un sous-domaine de la linguistique informatique qui travaille à la théorie et pratique de l'utilisation de l'outil informatique, uniquement, pour la traduction des textes écrits et oraux d'une langue naturelle à une autre.

On l'appelle également traduction assistée par ordinateur (TAO) ou traduction logicielle. C'est un processus basé sur des faits de substitution à partir de systèmes de corpus, de typologie, de reconnaissance d'énoncés ou de reconnaissance vocale, de rapports entre expressions idiomatiques, proverbes, termes et expressions jargonneux,

[2] LAROUSSE-SELECTION. Nouveau petit Larousse en couleurs. Paris : édition 1968.
[3] Paul ROBERT. Le petit Robert. Paris : édition 1981.

## II.4 Les paradigmes de la traduction automatique [7]

Les trois paradigmes essentiels autour desquels la traduction automatique s'est développée sont les suivants :

**II.4.1 La traduction symbolique :** Pour qu'un système puisse traduire un document en entrée et produire un document en sortie, il doit pouvoir utiliser les informations et les données nécessaires, en effet, en traduction symbolique, des experts encodent explicitement leurs connaissances.

**II.4.2 Le paradigme des règles :** Un trait caractéristique des systèmes basés sur les règles a été la transformation ou 'mappage' des représentations en forme d'arbres étiquetés. Ces systèmes formalisent des règles de réécritures et proposent une série de transformations des arbres: un arbre morphologique est transformé en un arbre syntaxique, un arbre syntaxique en un arbre sémantique, un arbre d'interface du texte source en un arbre équivalent du texte cible, etc.

Un arbre doit donc satisfaire des conditions précises:

1- Posséder une structure particulière et contenir des unités lexicales particulières ou des traits syntactiques ou sémantiques particuliers. En plus, les arbres eux-mêmes sont testés par les règles de formation et de transformation.
2- Etre conforme aux règles grammaticales du niveau en question: morphologie, syntaxe, sémantique, etc. Les grammaires et les règles de transformation déterminent les conditions ou contraintes qui limitent les possibilités de transfert d'un niveau à un autre et, en somme, d'un texte de la langue source à un texte de la langue cible.

**II.4.3 Le paradigme statistique :** Le paradigme statistique repose sur l'exploitation de corpus de traduction qui permettent de déterminer la traduction la plus fréquemment utilisée, par les spécialistes, pour une expression donnée.

Cette technique a été vulgarisée par *Google Translate* :

> « Notre système adopte une méthode différente: d'une part, nous introduisons dans l'ordinateur des milliards de mots provenant de textes monolingues dans la langue cible; d'autre part nous ajoutons des textes mettant en parallèle les deux langues. Ces derniers sont créés à partir d'échantillons de traductions réalisées par des traducteurs professionnels. Nous appliquons ensuite des techniques d'apprentissage statistique pour créer un modèle de traduction. Nous avons obtenu d'excellents résultats dans le cadre d'évaluations réalisées dans ce domaine. »[4]

Pour répondre aux insuffisances que présente de plus en plus la méthode des règles, en raison de la complexité du comportement des langues naturelles, certains laboratoires et groupes de recherche s'orientent vers les méthodes statistiques. Les ordinateurs devenant

[4] http://www.google.fr/intl/fr/help/fac_translation.html#statmt

de plus en plus puissants, il est désormais possible de puiser dans les immenses corpus des bases de données informatisées pour réutiliser des fragments de phrases déjà traduites par les traducteurs professionnels. C'est la naissance de la méthode de traduction statistique.
Elle est initiée par IBM, mais atteint son apogée dans les années 2000 avec Google qui récupère toutes les traductions existantes sur Internet pour construire l'architecture de son outil de traduction *Google Translate*. Cette approche statistique s'appuie sur des corpus bilingues alignés. En effet, un lien est créé entre chaque partie du texte de la langue source et la partie correspondante dans la langue cible. Ce lien est généralement créé au niveau de la phrase.

Enfin, l'approche statistique nécessite la définition d'un modèle de traduction qui puisse calculer efficacement les probabilités de traduction entre les mots, les suites de mots et les autres constituants de la phrase.

**II.4.4 Les tendances actuelles :** Pour des résultats plus efficaces, les logiciels de traduction automatique pratiquent de plus en plus des méthodes d'intégration de différents paradigmes. Ils combinent les paradigmes pour bénéficier des forces de chacun dans un système bien construit: c'est le modèle hybride. Par exemple : *Systran*, *Google Translate* et *Bing Translator* qui opèrent un système hybride.
Systran exploite un moteur de traduction hybride qui intègre une analyse statistique à l'analyse sémantico-syntaxique traditionnelle du texte source. Cette approche permet au logiciel de choisir la solution la plus fréquente entre deux propositions du moteur sémantico-syntaxique.

Ce moteur hybride permet à Systran de se positionner en leader du marché, à ce jour. Auparavant, la méthode utilisée par les logiciels était fondée sur un système d'analyse sémantico-syntaxique. Le moteur analysait chaque phrase source et créait l'arbre syntaxique permettant de représenter ses composantes et les relations qui les unissent. Puis, chaque expression était traduite en faisant appel à un dictionnaire. Une fois l'arbre entièrement traduit, le logiciel restituait la phrase cible. Le dictionnaire constitue alors un élément central: plus il est complet, meilleur est le résultat. Pourtant, même avec des dictionnaires très fournis, il est presque impossible de produire une phrase cible totalement correcte dans la mesure où le dictionnaire, qui est un recueil de données lexicales aura du mal à rendre compte des mots et expressions contextualisés ou nouveaux.

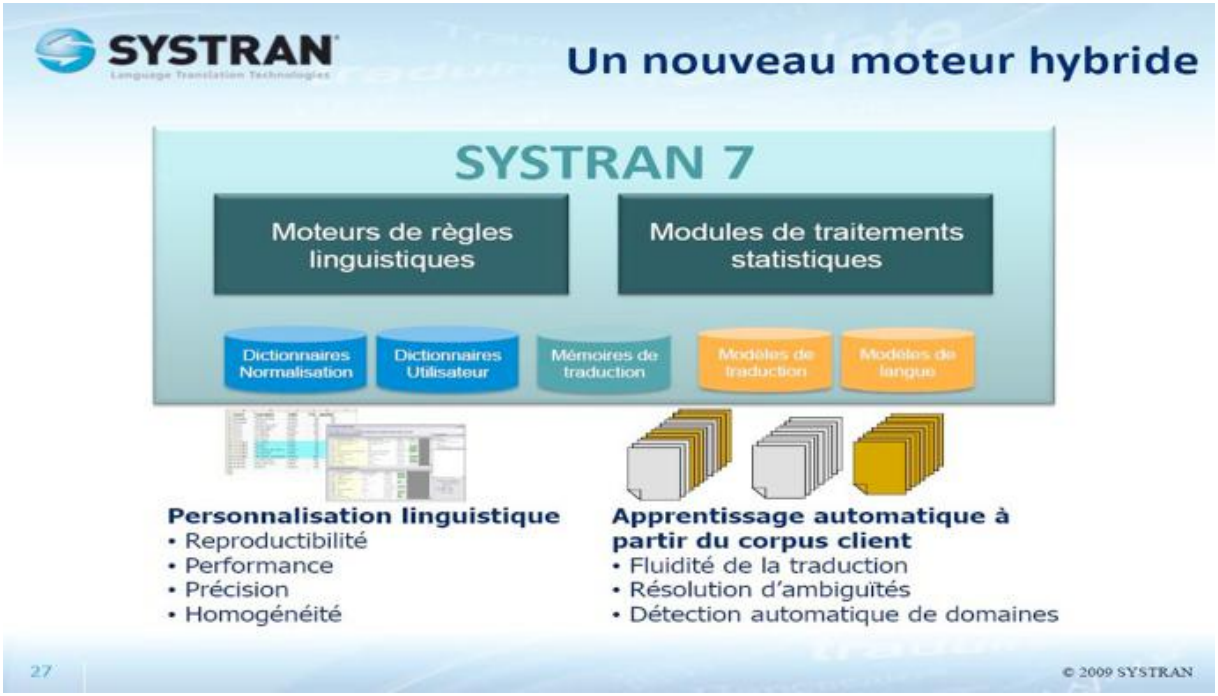


**Figure II.4.4: La base architecturale de *Systran 7* [7]**

## II.5 Les fonctions de la traduction automatique [8]

Selon John Hutchins (Hutchins, 2004 : 13-18), quatre fonctions particulières sont assignées à la traduction automatique :

**II.5.1 Fonction de dissémination :** consiste à produire un brouillon traduit du texte qui devra par la suite être post-édité manuellement pour aboutir à une traduction correcte.

**II.5.2 Fonction d'assimilation** : consiste à extraire des informations à partir du texte traduit automatiquement, sans considération pour la qualité du texte cible.

**II.5.3 Fonction d'échange :** consiste à utiliser la traduction automatique comme « interprète » de textes électroniques devant être traduits simultanément, comme par exemple les « chats », les pages web ou les courriers électroniques rédigés dans une langue étrangère.

**II.5.4 Fonction d'outil d'accès à l'information en langue étrangère** : ceci par l'interrogation d'un système de base de données. L'accès à une base de données par l'intermédiaire d'un logiciel de traduction automatique permet de recueillir des informations non-textuelles, comme des images.

En dehors de la fonction de dissémination, et plus rarement de la fonction d'assimilation, le recours au traducteur humain n'est pas prévu pour le genre de tâches assignées à la traduction automatique.

Il faut donc se poser la question de savoir si une post-édition d'un document traduit automatiquement peut présenter de l'intérêt par rapport à une traduction par écrasement comme elle est pratiquée par la plupart des traducteurs.

Pour illustrer ses propos, l'auteur a soumis tous les exemples authentiques ou construits recensés dans cette contribution à une traduction automatique à l'aide de la dernière version (payante) du logiciel de traduction automatique *Systran V6 Premium Translator* fonctionnant sur le modèle du transfert ainsi qu'avec l'outil de traduction (gratuit) de Google, *Google translate*. Pour simplifier, alors que les systèmes par transfert analysent le texte en langue source, en transfèrent les éléments lexico-syntaxiques dans la langue cible pour générer un texte en langue cible sur la base d'un modèle de langue complexe, les systèmes statistiques puisent à l'aide de modèles mathématiques compliqués dans d'immenses corpus parallèles des portions de textes déjà traduits pour les réassembler dans des phrases en langue cible.

D'un point de vue purement linguistique, le modèle de langue est beaucoup plus élégant, il est toutefois nettement plus difficile à mettre en œuvre du fait que la langue a tellement d'irrégularités et d'idiosyncrasies, que les formaliser toutes, semble illusoire. Le modèle statistique s'affranchit sinon totalement, du moins en grande partie d'une analyse linguistique. Comme pour les mémoires de traduction, il s'agit de piocher des séquences de textes déjà traduites, l'art résidant dans l'assemblage et la construction d'un texte entier, ce que ne fait pas la mémoire de traduction. Le modèle statistique s'est donc presque affranchi du linguiste et demeure la chasse gardée des informaticiens et des mathématiciens. Bien entendu, comme pour les mémoires de traduction, pour que le modèle fonctionne bien, il faut que les corpus soient à la fois nombreux et de bonne qualité.

## II.6 Logiciel de traduction [3]

Dans cette section on définira le logiciel de traitement, et on citera également ses composants :

### II.6.1 Définition

Un logiciel de traduction peut être définie comme étant une application ou un outil du traitement automatique du langage qui vise à traduire un document en entrée (lettres, rapports, articles, site web ...) d'une langue (source) vers une autre langue (cible).
Jusqu'à présent, les logiciels de traduction peuvent fonctionner en environnement PC , Internet, Intranet soit sous la forme d'une application autonome (logiciels de traduction par exemple), soit sous la forme de fonctions de traduction directement intégrées dans une application (dans un intranet , dans des moteurs de recherches, comme Google, dans les barres d'outils de certains logiciels qu'on peut ajouter à l'ensemble des barres d'outils qu'on installe dans un ordinateur, ou encore dans les logiciels de traitement de texte Word ou Excel...).
Il existe par ailleurs d'autres outils d'aide à la traduction qui sont complémentaires à un logiciel de traduction ce sont les mémoires de traduction comme (Trados) et dictionnaire électronique bilingues comme (Collins) ou multilingues comme (EuroDico).

Le souci majeur des spécialistes, des groupes de recherche et développeurs; est de faire en sorte que le texte traduit doit être autant que possible cohérent dans la langue cible et restituer l'information contenue dans le texte d'origine.

## II.6.2 Composants d'un logiciel de traduction

Généralement, un logiciel de traduction est constitué de :

**II.6.2.1 Les règles linguistiques :** Le moteur de traduction est le cœur technologique du logiciel de traduction. C'est là où se réalise concrètement la traduction automatique. On y applique les règles linguistiques de transformation propre à chaque langue.

**II.6.2.2 Dictionnaires :** Les dictionnaires intégrés aux logiciels de traduction, ne doivent pas être comme dans les premières générations de traducteurs automatiques, de simples dictionnaires de mots faisant uniquement office d'une liste mots ou d'expressions avec leurs traductions, chaque construction linguistique doit être définie avec des informations linguistiques précises (syntaxe, sémantique, pragmatique…) dans la langue source puis dans la langue cible. Ces informations sont ensuite gérées par le moteur de traduction. Plus les dictionnaires sont riches, plus la traduction obtenue est précise.

**II.6.2.3 L'interface :** L'interface des logiciels de traduction a un rôle important, puisque c'est elle qui permet la réalisation et l'exploitation de la traduction. En fait une fois le texte dans la langue cible est obtenu, l'utilisateur souhaitera le relire, le vérifier et le personnaliser pour l'adapter à son style, à son activité, à ses besoins, ceci dépend de l'utilisation finale de la traduction (compréhension, diffusion, publication externe…) et s'effectue via l'interface.

L'interface d'un logiciel de traduction se compose de :

**II.6.2.3.1 Module de traduction :** Les modules de traduction peuvent être définis comme étant l'ensemble des éléments, des options et des fonctionnalités qui constituent d'un point de vue général le logiciel de traduction. on cite ici la :

1- Présentation du texte source et de sa traduction sous forme de deux fenêtres l'une en dessus de l'autre ou l'une à côté de l'autre.
2- Conservation de la mise en page, option que ne fournissent pas l'ensemble des traducteurs.
3- Intégration dans des logiciels Microsoft Word (95,98,2000) dans les barres d'outils d'Internet Explorer (4.5) ou des moteurs de recherche.
4- Traduction de documents en intégral ou encore de sites web en partie ou complètement.

**II.6.2.3.2 Outils de traduction et de révision :** Comme leur nom l'indique, les outils de traduction et de révision sont l'ensemble des fonctionnalités permettant une relative amélioration du résultat de la traduction automatique, il s'agit de :

1- L'alignement des paragraphes source et cible,
2- La liste et marquage des mots inconnus.
3- La création et marquage des mots réservés (mots ou parties de textes que l'on ne souhaite pas traduire)
4- La proposition d'alternatives de traduction quand le contexte est ambigüe.
5- L'enregistrement du texte en mode bilingue.
6- La possibilité de traduction paragraphe par paragraphe ou texte entier.

## II.7 Fonctionnement de la traduction automatique [3]

La traduction automatique se heurte en fait à des difficultés aussi bien d'ordre lexical et syntaxique que sémantique et pragmatique.

C'est pour cela qu'afin de fonctionner convenablement et obtenir les meilleurs résultats possibles ; durant la phase de programmation, en plus des règles fondamentales de grammaire de la langue source et de la langue cible ainsi que des exceptions à ces règles, tout vocabulaire doit être accompagné, des dictionnaires des traducteurs automatiques.

### II.7.1 Le principe de fonctionnement de la traduction automatique [3]

Lors de la traduction, le moteur de traduction procède selon les étapes suivantes :

1- **Lecture** : à l'issue de cette étape, tous les mots ont été délimités c'est-à-dire divisés en unités à traduire séparément.
2- **Consultation des dictionnaires :** au terme de cette opération, les mots doivent être reconnus. Dès lors, l'ensemble des informations reprises dans les entrées des dictionnaires doivent être accessibles.
3- **Analyse :** cette opération est divisée en deux parties successives. Au terme de la première, la phrase est analysée dans sa totalité, mais certaines ambigüités subsistent. La seconde étape vise à analyser la phrase plus complètement. A l'issue de cette opération, la phrase est prête à être traduite.
4- **Transfert :** les programmes visent à donner une première traduction. Au terme du processus, seules les ambigüités sémantiques, provoquées par exemple par des cas de polysémie et des empilages, subsistent.
5- **Synthèse :** A l'issue de cette étape, la traduction définitive est en place. Le texte est compris entièrement et toutes les ambigüités sont résolues.

6- **Impression :** après l'impression, la traduction finie est prête à la révision ou à la publication.

La figure suivante explique de nouveau le processus de traduction (les éléments du système sont repris selon l'ordre chronologique dans lequel ils interviennent)

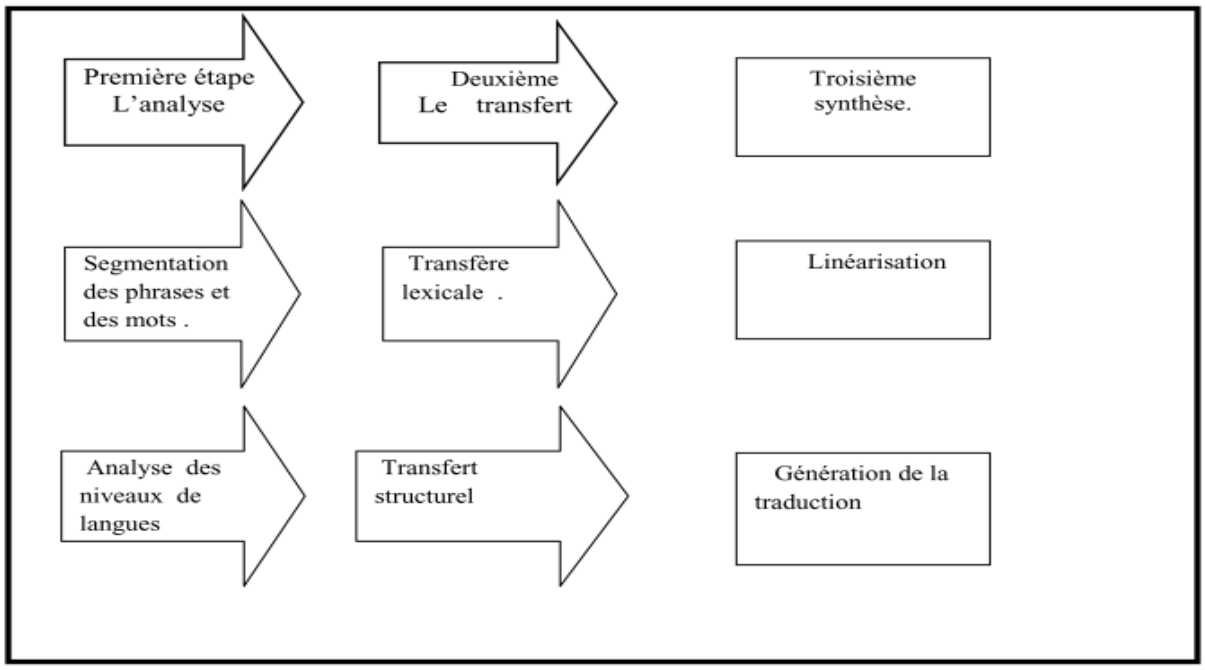


**Figure II.7.1 : Schéma récapitulatif du processus de traduction automatique [3]**

## II.8 Les treize péchés capitaux de la TA [8]

Dans un article déjà très ancien, Anne-Marie Loffler-Laurian (Loffler-Laurian, 1983 : 65-78) relève douze catégories d'erreurs générées par le système Systran. Plus d'un quart de siècle plus tard, les mêmes erreurs sont retrouvées, malgré les progrès qu'ont faits les systèmes de traduction automatique. L'autre a relevé dans son article treize types d'erreurs générés par les systèmes de traduction automatique.

**II.8.1 Premier problème, polysémie et homonymie** : La polysémie constitue le problème le plus souvent signalé en matière de traduction automatique. Les mots ne fonctionnent pas tant comme des unités discrètes, c'est-à-dire bien délimitées et séparées les unes des autres, mais comme des occurrences sujettes à un certain nombre de variations sémantiques au sein d'un champ.

La polysémie, pluralité de significations au sein d'un continuum sémantique pour un même mot, pose problème également au traducteur humain pour qui il est parfois difficile de trouver la nuance exacte. Tantôt la polysémie se retrouve d'une langue à l'autre (exemple 1), tantôt pas (exemple 2).

***Exemple 1 : la traduction de l'adjectif polysémique « libre »***

La voie est libre (non encombrée)

L'entrée est libre (gratuite)

Le prisonnier est libre (n'est plus en captivité)

The way is free (not encumbered)

The entry is free (free)

The prisoner is free (is not any more in captivity) [Systran]

***Exemple 2 : la traduction de l'adjectif polysémique « léger »***

Ce sac à dos est léger. (a peu de poids)

Le directeur est léger dans son travail. (est negligent)

This backpack is light. [Google]

The director is light in his work. [Google]

L'homonymie qui concerne des mots de même graphie et de sens radicalement différents pose également des problèmes :

L'un des problèmes de la traduction automatique est qu'en règle générale, très peu de variantes lexicales sont présentes. C'est un problème de dictionnaire électronique qui peut être résolu et certains programmes, comme *Reverso Pro*, proposent des variantes de traduction dans la mesure où une unité comporte des homonymes.

**II.8.2 Deuxième problème, l'ambiguïté syntaxique** : L'ambiguïté syntaxique est le second problème couramment évoqué, il apparaît du fait que certaines structures syntaxiques ne sont pas claires sans connaissance du monde :

***Exemple 3: to fly gliders and *to clean fluids***

Cleaning fluids can be dangerous (cleaning fluids non pas *to clean fluids)

Flying gliders can be dangerous (double interpretation: flying gliders et to fly gliders)

*To clean fluids can be dangerous

To fly gliders can be dangerous

En anglais, les verbes « to fly » et « to clean » sont transitifs. Il existe cependant une restriction quant aux arguments ayant la fonction d'objet direct. Ainsi « to fly » demande comme objet un « objet volant ». Bien que moins sélectif quant à la nature de l'objet, « to clean » est incompatible avec « fluids ». L'ambiguïté syntaxique fait appel

au contexte argumental et ne peut être résolue que par la prise en compte du niveau lexico-syntaxique (Gross, 1995 :16-19).

**II.8.3 Troisième problème de la TA, l'ambiguïté référentielle** : La question de la référence (rapport entre le texte et la part non linguistique de la pratique où il est produit et interprété) est une question d'interprétation qui suppose par définition une interprétation cognitive.

Les pronoms réfèrent ainsi à certains mots ou antécédents qui ne sont pas toujours clairs sans connaissance du monde.

La traduction automatique est effectuée phrase par phrase et les référents peuvent se situer hors d'une phrase particulière.

***Exemple 4 : le pronom « le » et son référent***

John a heurté le vase du pied et l'a cassé. (le vase ou le pied?)

John ran up against the vase of the foot and broke it. [Systran]

John struck the foot of the vase and broke. [Google]

L'ambiguïté référentielle fait aussi appel à la connaissance du monde plus que du contexte et dans l'état actuel de la TA, ne peut être résolue de façon satisfaisante.

**II.8.4 Quatrième problème, les expressions floues (*fuzzy hedges*)** :Ce sont des mots ou groupes de mots au caractère idiomatique marqué, donc très dépendants de l'organisation sémantique de la langue source, qui sont difficiles à traduire et dont le rôle est d'exprimer une approximation – "words whose job it is to make things more or less fuzzy" selon Lakoff (Lakoff, 1972 : 183). À titre d'exemple on relèvera « en fait », « d'ailleurs », « en un sens » en français, « somehow », « a sort of », « actually » en anglais, etc.

***Exemple 5 : traduction de l'expression floue « en un sens » / « in a certain way »***

Parler n'est-il pas toujours en un sens donner sa parole ?

To speak isn't always in a direction to give its word? [Systran]

Talking is not always in a sense to give his speech? [Google]

La question de la traduction des termes flous constitue un problème de lexique qui se situe souvent au niveau polylexical, il n'est pas possible de le résoudre sans prendre en compte le contexte adjacent. Ces termes flous posent problème aux systèmes par transfert et statistiques du fait d'un contexte d'apparition très variable.

**II.8.5 Cinquième problème, idiotismes et métaphores** : Les idiotismes ou expressions idiomatiques ou encore phrasèmes chez Mel'čuk (Mel'čuk, 1998), ainsi que les

métaphores revêtent une coloration culturelle marquée qu'il est difficile de traduire mot à mot.

***Exemple 6 : traduction de l'expression idiomatique « à couteaux tirés »***

Nicolas Sarkozy et Jean-François Copé sont désormais à couteaux tirés. (20minutes.fr)

Nicolas Sarkozy and Jean-François Cope are from now on with drawn knives. [Systran]

Nicolas Sarkozy and Jean-François Copé are now at loggerheads. [Google]

Nicolas Sarkozy und Jean-François Copé sind von nun an an gezogenen Messern. [Systran]

Nicolas Sarkozy und Jean-François Copé sind nun Messer aus. [Google]

La traduction anglaise de Google « to be at loggerheads (être en désaccord) », bien que moins précise, est acceptable.

Du fait qu'ils fonctionnent à partir de bases de données de textes traduits humainement, les systèmes statistiques de traduction automatique peuvent se révéler plus performants que les systèmes par transfert pour ce genre de problèmes.

**II.8.6 Sixième problème, la néologie** : La langue générale et plus encore la terminologie évoluent et les logiciels de traduction automatique n'incluent pas toujours les dernières évolutions lexicales.

***Exemple7 : traduction des néologismes « internautes » et « Web star »***

Ancienne comédienne, Luna Sentz met son talent au service des internautes en animant des émissions interactives en direct sur le site de Canal+. Une Web star est née. (L'Ordinateur Individuel)

Former actress, Luna Sentz puts her talent at the service of the Net surfers by animating interactive emissions on line on the site of Canal+. A Web star was born. [Systran]

Former actress, Luna Sentz puts his talent to the Internet in facilitating interactive programs live on the site of Canal +. Web is a star born. [Google]

La néologie suppose une actualisation régulière des dictionnaires électroniques, avec des équipes de lexicographes qui travaillent en arrière-plan pour les systèmes par transfert. Grâce à ses immenses corpus de textes traduits relatifs aux nouvelles technologies, Google s'en sort ici particulièrement bien.

**II.8.7 Septième problème, les noms propres:** Le problème des noms propres est sans doute l'un des plus difficiles à résoudre en traduction automatique : d'une part, leur nombre est tellement élevé qu'un recensement exhaustif paraît pratiquement impossible. En effet, si l'on considère l'ensemble des noms de personnes, des noms de lieux, des

noms de marques, d'associations, d'organismes à l'échelle de la planète, on dépasse de loin pour une langue les dictionnaires de langue générale. A la difficulté du recensement des noms propres vient s'ajouter celle de leur orthographe, souvent fluctuante lorsqu'il s'agit de translittération ou de transcription d'une langue à l'alphabet non latin.

***Exemple 8 : un illustre inconnu, le Dr Michel Maure***

Un mandat d'arrêt a été délivré à l'encontre du Dr Michel Maure, 59 ans, auteur de multiples opérations de chirurgie esthétique ratées. (europe1.fr)

A warrant for arrest was delivered against Dr. Michel Moor, 59 years, author of multiple missed operations of cosmetic surgery. [Systran]

An arrest warrant was issued against Dr Michel Maure, 59, author of multiple cosmetic surgery operations failed. [Google]

La présence de noms propres vient souvent complètement bouleverser la traduction, le système ne parvenant plus à analyser la phrase de manière correcte.

**II.8.8 Huitième problème, les mots d'origine étrangère et les emprunts** : Les mots d'origine étrangère sont extrêmement fréquents dans la langue allemande et proviennent généralement de l'anglais ou du français. Ces mots ne sont généralement pas présents dans les dictionnaires électroniques du système, d'où un net avantage aux systèmes statistiques.

De plus en plus, on constate parallèlement à la mondialisation, une tendance dans la langue journalistique à la généralisation des emprunts de mots d'origine étrangère ou à la traduction mot à mot de certaines expressions qui prennent aussi racine dans la langue cible. L'expression anglaise « nothing in the pipeline », est devenue en français « rien dans les tuyaux ». Bien entendu, la langue d'emprunt est en général l'anglais pour ces expressions, la langue du « business » international.

***Exemple 9: traduction de l'expression idiomatique empruntée de l'anglais « rien dans les tuyaux » et « nothing in the pipeline »***

Most software companies are one product companies, and have nothing in the pipeline apart from upgrades. (techuser.net)

La plupart des fournisseurs de logiciel sont des compagnies d'un produit, et n'ont rien dans la canalisation indépendamment des mises à niveau. [Systran]

La plupart des entreprises de logiciels sont l'un des produits des entreprises, et n'ont rien en dehors de la canalisation de mise à niveau. [Google]

L'expression idiomatique est traduite dans la langue cible comme s'il s'agissait d'une proposition libre.

Christine Lagarde, fait valoir de son côté qu' « il n'y a, à ce jour, strictement rien dans les tuyaux ». (tradingsat.com)

Christine Lagarde, puts forward on her side that “there is not, to date, strictly nothing in the pipes”. [Systran]

Christine Lagarde, argues in turn that "there has, to date, nothing in the pipes." [Google]

Certains emprunts de l’anglais jouissent d’une grande popularité, surtout dans la langue des affaires.

***Exemple 10: traduction de l’emprunt « business model »***

Les pirates sont innovants, ils mettent en évidence les problèmes du marché et montrent la voie à de nouveaux business models. (ecrans.fr)

The pirates are innovating, they highlight the problems of the market and show the way with new businesses models. [Systran]

The pirates are innovative, they highlight the problems of the market and show the way to new business models. [Google]

Là encore, il s'agit d'un problème de qualité des dictionnaires électroniques qui peut être résolu dans un système par transfert par création d’un dictionnaire des emprunts. Le système statistique se montre un peu plus performant du fait qu’il se fonde sur des corpus de textes traduits humainement.

**II.8.9 Neuvième problème, les séparateurs**: Les signes de ponctuation ainsi que certaines abréviations posent problème aux systèmes de traduction automatique. Le fait que le point n’ait pas toujours une fonction de séparateur de phrases constitue un phénomène bien connu en matière de segmentation en français

***Exemple 11: séparateur et adjectif numéral ordinal***

Montag, den 18. August 2008

Lundi le 18 août 2008 [Systran]

Lundi, le 18 Août 2008 [Google]

Ce problème des séparateurs peut tout à fait être résolu dans la plupart des cas, ce qui implique des modules supplémentaires dans le moteur de traduction.

Les sigles ne prennent en principe plus de points entre les différentes lettres qui les composent, ce qui constitue une erreur potentielle de moins pour la question des séparateurs.

**II.8.10 Dixième problème, les sigles et les acronymes** : Les sigles (épelés), séparés ou non par des points, ainsi que les acronymes (prononcés comme une unité phonique et donc sans points séparant les différentes lettres) sont couramment employés dans les textes journalistiques. Certains se traduisent, d'autres pas. Certaines langues, comme l'allemand, utilisent aussi les acronymes anglais. Signalons toutefois que les acronymes s'écrivent le plus souvent en capitales et sans points abréviatifs : *UNESCO, ONU, OTAN, NASA*. Parfois aussi, on les écrit aussi comme des noms propres, avec une majuscule initiale : *Onu, Insee*. Les sigles perdent plus difficilement leurs points étant donné qu'ils se prononcent lettre après lettre : *S.N.C.F.*

***Exemple 12 : traduction du sigle d'une organisation internationale connue***

L'Organisation mondiale du commerce (OMC) est la seule organisation internationale qui s'occupe des règles régissant le commerce entre les pays. (wto.org)

The World Trade Organization (WTO) is the only international organization who deals with the rules governing the trade between the countries. [Systran]

The World Trade Organization (WTO) is the only international organization dealing with the rules governing trade between countries. [Google]

Système par transfert et système statistique viennent à bout de ce genre de problèmes. Un système comme *Google* a naturellement à sa disposition les pages traduites desdites organisations, d'où la qualité de la traduction automatique réalisée sur la base de corpus parallèles.

**II.8.11 Onzième problème, les synonymes**: La question de la synonymie est l'une des plus cruciales en traduction car elle traduit la richesse lexicale d'une langue et la compétence d'un traducteur. De nombreux mots ne se différencient les uns des autres que par des différences, parfois infimes, mais nécessaires pour reproduire telle ou telle nuance de style ou de sens dans tel ou tel contexte. Un logiciel de traduction ne dispose généralement que d'un nombre limité de variantes pour traduire telle ou telle unité. La traduction peut ainsi apparaître compréhensible, mais peu élégante, voire maladroite.

***Exemple 13: traduction de « banner » par « bannière » au lieu de « banderole »***

Two British Free Tibet campaigners are in custody in China after unfurling a Tibetan flag and banner outside the Olympic stadium. (freetibet.net)

Deux militants libres britanniques du Thibet sont dans la garde en Chine après unfurling un drapeau et une bannière tibétains en dehors du stade olympique. [Systran]

Deux British Free Tibet militants sont en garde à vue après le déploiement en Chine, le drapeau tibétain et la bannière à l'extérieur du stade olympique. [Google]

Quasi-synonymes, « bannière » = « étendard d'une confrérie, d'une société » ne s'en distingue pas moins de « banderole » = « grande bande de tissu qui porte une inscription (en signe de protestation) ».

**II.8.12 Douzième problème, la transposition**: Au sens classique, la transposition en traduction consiste à traduire une unité lexicale d'une classe (nom, verbe, adjectif, adverbe) par une unité lexicale d'une autre classe. La transposition est assez fréquente lorsqu'on traduit des langues romanes vers les langues germaniques, les premières ayant souvent recours à des nominalisations ou les secondes préféreront des expressions verbales.

***Exemple 14 : transposition de « house for sale » en « maison à vendre »***

Detroit has a bunch of run down houses for sale in the $ 30,000 range.

Detroit a un groupe de maisons de course vers le bas à vendre dans la gamme $30000. [Systran]

Detroit a un tas de courir les maisons en vente dans la gamme $ 30,000. [Google]

L'expression « house for sale » confine à l'idiotisme et les deux systèmes ont procédé à la dislocation de l'expression qui devient incompréhensible. Mais le problème consistait à ne pas traduire « for sale » par une suite « préposition + nom » mais « préposition + verbe ». C'est une gageure pratiquement impossible à résoudre pour un système par transfert et là encore, même s'il ne brille pas, le système statistique se révèle meilleur. Paradoxalement, les systèmes s'en sortent mieux à la transposition de « maisons à vendre » du français vers l'anglais.

***Exemple 15 : transposition de « maison à vendre » en « house for sale »***

Dans tout le pays on organise des foreclosure tours, visites organisées de maisons à vendre.

In all the country one organizes foreclosure turns, visits organized of houses for sale. [Systran]

Across the country on organizes foreclosure tours, tours of homes for sale. [Google]

**I.8.13 Treizième problème, l'orthographe**: L'orthographe, lorsqu'elle est défectueuse, est un ennemi de la traduction automatique, c'est une remarque triviale. Les systèmes de traduction travaillant sur du texte analysent tout mot mal orthographié comme mot inconnu et tout mot inconnu ne peut avoir de traduction et est donc laissé tel quel. Un mot inconnu ne peut être analysé, c'est-à-dire rattaché à une classe de mots comme les noms, les verbes, etc. De ce fait, un mot inconnu provoque systématiquement des erreurs d'analyse dans un système par transfert. L'étape qui doit précéder toute traduction automatique consiste donc en une pré-édition dont le but minimal sera de corriger les fautes d'orthographe.

***Exemple 16 : traduction d'une phrase mal orthographiée***

Alors qu'un traductteur humain est cpable de traduire une phrase mâle orthographiée, il en va tout ôtrement d'un logiciel.

Whereas a human traductteur is cpable to translate a spelled male sentence, it goes from there all ôtrement a software. [Systran]

While a traductteur human cpable to translate a sentence spelled male, that's a ôtrement of software. [Google]

Si certains problèmes relatifs à la TA (ambiguïté, transposition) sont imputables à la question de la « connaissance du monde », la plupart d'entre eux proviennent d'un codage insuffisant des dictionnaires ne prenant pas encore en compte toutes les avancées de la linguistique dans les systèmes par transfert. L'avenir semble donc appartenir aux systèmes statistiques disposant de corpus parallèles les plus larges possibles. Quoiqu'il en soit, la traduction automatique connaît, avec de tels systèmes dérivés des mémoires de traduction, un renouveau dont il est impossible de ne pas tenir compte. Ils remettent aussi au goût du jour la technique de la post-édition, c'est-à-dire la correction d'une ébauche traduite automatiquement.

## II.9 Conclusion

Nous concluons ce chapitre en soulignant que la traduction automatique est historiquement et concrètement la première application du traitement automatique du langage, qu'elle est donc dans la poursuite et à l'instar de toutes les applications de ce dernier, et le résultat de collaboration entre différents spécialistes. Néanmoins après soixante ans d'existence, elle présente encore des lacunes et rencontre des problèmes divers.

# *Chapitre III*

## *Mots composés et terminologie Informatique*

# Chapitre III: Mots composés et terminologie informatique

## III.1 Introduction

Une langue subit en continu des modifications et changements. De jour en jour des mots meurent, d'autres apparaissent, le lexique évolue. Pour chaque mot chaque concept et chaque réalité, un mot.

Les données diffèrent selon les domaines, qu'ils soient techniques ou scientifiques, les professionnels emploient des termes précis. Un terme est en effet le principe d'élaboration de chaque terminologie.

Dans ce chapitre on s'intéressera à la terminologie informatique, mais d'abord, nous avons été amenées à apporter quelques définitions nécessaires à la compréhension des sections suivantes.

## III.2 Définitions [9] [10]

### III.2.1 Qu'est ce que la terminologie ?

Le mot terminologie signifie un « ensembles de mots techniques appartenant à une science, un art, un auteur ou un groupe social », par exemple, la terminologie de la médecine.

### III.2.2 Mot

Un « mot » est un son ou groupe de sons articulés, constituant une unité porteuse de signification à laquelle est liée, dans une langue donnée, pour désigner un être, un objet, un concept, etc.

### III.2.3« Terme »

Est la désignation d'une notion sous forme de lettres, de chiffres, de pictogrammes ou d'une combinaison quelconque de ces éléments ISO (1990).

### III.2.4 forme simple

Une forme simple est une séquence consécutive non vide de caractères apparaissant entre deux séparateurs. Cette définition est orthographique, par exemple les mots assietes, bele sont des formes simples même s'ils n'existent pas dans la langue française.

### III.2.5 Mot simple

Un mot simple de la langue est une forme simple qui constitue une entrée d'un dictionnaire des mots simples fléchis de cette langue.

Exemple :

-chaise,
-chien,
-fruit,
-écrire.

### III.2.6 forme composée

Une forme composée est une séquence consécutive d'au moins deux formes simples de séparateurs.

### III.2.7 Mot composé

Un « mot composé » est une forme composée qui constitue une entrée du dictionnaire des mots composés fléchis de cette langue.
Exemple :

-Sac à dos,
-Avion de chasse,
-porte-clés.

### III.2.8 Synapsie

Dans des articles postérieurs de quelques années à ceux de Pottier, Benveniste (1966,1967) identifie des synapsies, qu'il définit comme des groupes de lexèmes formant une unité nouvelle à signifié unique et constant.
On trouve le noyau initial dans des exemples déjà anciens comme : « voiture de course », « robe de chambre », « clair de lune ». La structure formelle de ces synapsies (Nom + Adjectif, ou Nom relié à un autre Nom par les prépositions *à* ou *de),* qui entraîne leur ressemblance avec des syntagmes nominaux « ordinaires » masque parfois leur singularité sémantique ; le sens exact (en physique) de synapsies comme *trou noir* (et *trou blanc* créé par analogie avec le premier) est certainement fort peu connu en dehors des milieux spécialistes, bien que l'expression « dise quelque chose » en soi à la plupart des francophones. Ce qui contribue, précisément, à définir ces synapsies (relativement aux syntagmes nominaux), c'est que leur sens n'est pas produit par composition des sens des éléments qui les constituent ; il s'opère une sélection réciproque parmi les traits sémantiques des lexèmes de base, à partir de laquelle se dégage un sens, au moins partiellement arbitraire, et largement imprévisible : cette sélection explique la (quasi) synonymie *de feuille de paye* et *bulletin de salaire,* alors que *paie* n'est pas synonyme de *salaire* (ils s'opposent même dans certains emplois), et encore moins *feuille* de *bulletin.*

Aujourd'hui ce type de composition prend une extension importante qu'il est appelé à une productivité indéfinie : il sera la formation de base dans les nomenclatures techniques.

## III.3 Notion de composition [10]

Le français, comme d'autres langues romanes, construit des noms, noms d'objets ou d'évènement, par composition.

Divers travaux de recensement des mots composés du français existent, citons les travaux de (Gross.M 1986a) sur les adjectifs et les adverbes composés. (Piot M. 1988) sur les conjonctions composées, (Danlos L. 1980) sur certains couples verbe-complément, (Gross M. 1988) sur les verbes composés et les phrases idiomatiques.

Un terme peut être simple s'il ne contient qu'un seul mot, ou composé s'il en contient plus d'un et qu'un mot composé est construit à partir de mots simples (Daille B.1994).

La notion de composition nominale, a été abordée dans divers travaux de (Monceaux A. 1993) et (Mathieu-colas M.1988) pour la syntaxe de noms composés.
La principale problématique associée aux noms composés est de pouvoir détecter si un groupe de mots constituent un terme pour la science ou discipline retenue.
Le problème est complexe pour le linguiste, aussi l'aide du spécialiste de la terminologie étudiée est souvent d'un grand secours.
Dans ce qui suit, on procédera à l'étude des mots composés, et la distinction entre groupe nominal libre et nom composé. On verra également les critères définis par Gaston Gross et ceux de Max SILBERZTEIN, qui permettent d'apporter une aide afin de distinguer un nom composé d'un groupe nominal libre.

## III.3.1 Mot composé [11]

Deux éléments ou plus, appelés bases, sont associés (composition = base + base). Ces éléments sont en principe des unités lexicales autonomes (contre - poids). Mais ils peuvent ne pas avoir d'existence autonome (camé-scope ; écho dans écho-graphie).

Les noms composés sont nombreux : grande surface, mode d'emploi, touche-à-tout.
Mais aussi :

- Les verbes: court-circuiter
- Les adjectifs : bleu clair, bon dernier
- Les déterminants : une tonne de, une dizaine de, beaucoup de, vingt-quatre
- Les adverbes : ci-devant, ci-dessus
- Les pronoms : celui-ci, le mien, lequel

### III.3.1.1 La composition proprement dite ou composition populaire

Il s'agit de tous les mots composés dont les éléments sont des mots français qui ont une existence autonome par ailleurs : portefeuille, porte -monnaie, chaise longue.
L'unité formée par composition portera une nouvelle signification qui dépasse celle de chaque élément pris isolément, comme par exemple : une chaise longue qui n'est pas littéralement une chaise qui est longue mais globalement, un fauteuil pliable destiné au repos en position allongée.

Les modes d'association possibles entre les constituants d'un composé, sont :

1- **La Soudure graphique des constituants :** portefeuille, clairvoyant. Ils tendent à être perçus comme des mots simples ;

2- **Les éléments sont liés par un trait d'union :** Chef-lieu, sourd-muet, cerf-volant (rappel : la réforme de l'orthographe de 1990, dans une volonté de simplification, préconise de remplacer le trait d'union par la soudure pour les composés en : contre-, intra -, entr(e)-, extra-, infra-, ultra-.
3- **Les éléments sont séparés par un blanc graphique :** bon sens, eau de vie, garde champêtre, à cause de, afin que.

### III.3.1.2 La composition savante

La plupart des termes scientifiques et techniques du français sont formés d'éléments savants grecs et latins. On parle dans ce cas de composition savante.
Les composés savants sont formés par la réunion de bases grecques ou latines, qui ne constituent généralement pas des unités lexicales autonomes (sauf abréviations : télé, auto).

Certains composés savants sont empruntés en bloc au latin ou au grec : aqueduc, aquaeductus, amphibie, amphibios. Mais la plupart des composés ont été forgés en français ; cependant, la voyelle de transition -o- relie en principe deux éléments de composition grecs (ex : baromètre) et la voyelle -i- deux éléments latins (ex: digitigrade).
Ces formations comportent le plus souvent des éléments de même origine.
- composés grecs : anthropologie, thalassothérapie, démocratie, topographie.
- composés latins : apiculture, multicolore, homicide, viticole.

Parfois on trouve des composés hybrides comme :
Génocide: Genos signifie «la race» en grec, caedere signifie «tuer» en latin ; ou base française et élément grec : hypotension, archiplein.

La composition savante peut être associée à la dérivation affixale. Ainsi, beaucoup de noms savants fonctionnent comme des bases pourvues du suffixe nominal –ie comme élément final : agronomie, radioscopie, télépathie.

Par ailleurs, certains éléments savants, par leur productivité et leurs combinaisons, se comportent comme des préfixes et tendent à être assimilés à des préfixes :
Anti-: antidote, antigel, hypo-: hypoallergénique, hypotension. Des éléments adverbiaux comme bien-, mal-, non -, arrière-(bienheureux, malformé, arrière -plan) ou prépositionnels comme après -, avant-, contre-, entre-, sur-, sous, outre-, sans-
(Service après-vente, avant-coureur, contre-épreuve, entre-deux, surdétermination, sous-louer, outrepasser).

## III.4 Degré de figement des noms composés [10]

La majorité des mots composés sont des noms composés. Dans (Gross M. 1990), ces critères permettant de calculer le degré de figement des noms composés de type « N de N » sont présentés. Les groupes « N de N » sont analysés dans (Gross G.1991).
Différents critères sont retenus pour décider si on a un nom composé que l'on pourra rajouter dans un dictionnaire ou un groupe nominal libre. Ils sont principalement fondés sur l'analyse des relations entre les deux noms. Ces critères sont les suivantes :

**III.4.1 « il n'y a pas de relation syntaxique entre les 2 noms » :** On n'a pas de figement si dans un groupe N de N, on peut établir une relation syntaxique entre les deux noms. Il y a deux cas :

**1- N1 est un concret :**

Exemples :
-l'ordinateur de John.
-John à un ordinateur.

-Un pied-de-biche.
-La biche a un pied.*, ici on a un nom composé, pied de biche

**2- N1 est un substantif prédicatif : « N1 de N2 » :**
Exemple :
Le voyage de John.

Diverses transformations sont possibles :
**-Nominalisation** : John a fait un voyage en Grèce ;
**-Relativation** : le voyage que John a fait en Grèce ;
**-Réduction du verbe support** : Le voyage de John en Grèce.

Dans ces cas, N2 est sujet du substantif prédicatif N1, dans les cas suivants, N2 peut représenter l'objet.

Exemple :
-L'aménagement du territoire. On a procédé à l'aménagement du territoire.
S'il y a une relation syntaxique entre les 2 substantifs d'un GN, chacun des éléments est susceptible de certaines transformations, dont la pronominalisation.

**III.4.2 Pronominalisation :** S'il y a une relation syntaxique entre les 2 substantifs d'un GN, chacun des éléments est susceptible de certaines transformations, dont la pronominalisation, nous avons deux cas, la pronominalisation de N2 et la pronominalisation de N1 :

**1-Pronominalisation de N2 :**
N2 peut être pronominalisé par un possessif
Exemples :
-L'ordinateur de John.
Son ordinateur
-Je prends la responsabilité de cet acte.
J'en prends la responsabilité.

**2-Pronominalisation de N1 :**
Dans les groupes nominaux libres, N1 peut être pronominalisé, particulier dans le cas de la coordination. L'absence de pronominalisation est un indice de figement.

-J'ai un cours d'art et de musique.
Mon cours d'art et de musique.

**III.4.3 Figement partiel :** Deux cas sont possibles : le figement partiel du 1$^{er}$ nom ou le figement partiel du 2$^{ème}$ nom.

**1-Figement partiel du 1$^{er}$ nom :**
Exemple :
-John a versé un nuage de lait.
-John a versé du lait

-*John a versé un nuage.

**2-Figement partiel du 2ème nom :**
Exemple :
-John a une fièvre de chacal.
-*cette fièvre est de chacal

## III.5 Les groupes nominaux productifs et les noms composés lexicalisés [10]

(Silberztein M.1993a) a défini des critères utilisés pour définir un nom composé. L'apport sémantique des composants du GN est examiné, dans le cas des mots composés généralisés (séquences figées) les constituants sont des entrées dans le dictionnaire des mots simples DELAS.
Les groupes nominaux libres sont des séquences analysables. Il existe une frontière entre noms composés et séquences libres. Dans (Silberztein 1993a), Silberztein M. définit quatre critères de construction des dictionnaires des noms composés qui sont les suivantes :

**III.5.1 L'atomicité sémantique :** Si l'on veut décider si un groupe de mots est un mot composé ou une séquence libre, il faut comparer les propriétés syntaxiques et sémantiques des séquences « candidates au statut de mot composé » aux propriétés générales des séquences libres.

Si toutes les propriétés syntaxiques et sémantiques d'une séquence peuvent être déduites des propriétés de ses constituants, on n'a pas besoin de les décrire dans un dictionnaire, la séquence est libre.

En revanche si au moins une propriété syntaxique ou sémantique ne peut pas être calculée, il faut associer explicitement cette séquence à ses propriétés et donc la placer dans le dictionnaire : la séquence est figée.

Exemples :
-Une voiture rapide.- une voiture qui est rapide.
-Une voiture électrique.-* une voiture qui est électrique.
On ne peut pas paraphraser avec la structure à verbe être.

### III.5.2 L'institutionnalisation de l'usage

**III.5.2.1 Termes institutionnalisés :** Ces termes sont institutionnalisés bien que semblant syntaxiquement et sémantiquement analysables.
Un groupe nominal plus ou moins figé (terme) désigne de nombreux éléments de notre environnement.

Exemples :
-Des chefs d'état ;
-un agent de sécurité ;
-Une machine à coudre :
-Une torche électronique.

On ne dira pas des « chefs de nation », des « chefs de pays », un « agent sécuritaire », une « machine de couture », ou une « torche à électricité ».

**III.5.2.2 Termes inexistants :** On ne peut désigner un certain nombre d'éléments de notre entourage que par des groupes nominaux libres ou des paraphrases, car les termes n'existent pas.

Exemples :
-Il a mis une chemise qui a un motif à voitures.* il a mis une chemise à voitures.
contrairement à :
-Il a mis une chemise qui a un motif à rayures, on peut également dire : il a mis une chemise à rayures.

**III.5.3 Restrictions distributionnelles :** Dans par exemple le dictionnaire DELAC, on pourra avoir des entrées telles que :
(NA) un produit (actif+ adhésif+agricole+alimentaire+apicole…)
Ces groupes nominaux seront reliés à d'autres groupes nominaux par des transformations. (Gross G.1991, Monceaux A. 1993,1997) :
(NA) (une production+exportation+excédent+….)adj= : une production agricole.
Afin d'effectuer ces transformations, des classes distributionnelles sont à construire, par exemple les classes :
-AdjProduit (adjectifs de produits tels que agricole, alimentaire, apicole, laitier..etc.),
-Nprof (noms de profession).
On se reportera à (Gross M.1986), pour plus de détails sur leur construction.

**III.5.4 Analyse transformationnelle :** par exemple avec les classes Na et NDN :
-un accord monétaire. –un accord sur des monnaies,
-un tube de verre. – un tube en verre.

## III.6 variantes terminologiques [10]

Plusieurs catégories de variantes existent (parmi elles, les abréviations et les sigles). L'auteur a considéré ici, pour les noms composés, les variantes terminologiques avec :

**III.6.1 la surcomposition :** Un composé se voit rajouter un modifieur ou une tête pour créer un terme plus complexe,

Exemple :
-Réseaux de microstations --> Connexion de [réseaux de microstations] :

**III.6.2 les insertions :** Un modifieur est inséré à l'intérieur d'un composé :
Insertion d'un adjectif :

Exemple :
-Réseau d'ordinateurs --> réseau [local] d'ordinateurs
-Insertion d'un nom :
Exemple :
-Protocole TCP/IP -> protocole [internet] TCP/IP

**III.6.3 la coordination :** Un terme est coordonné avec un autre terme par la mise en facteur de la partie commune. (Domingues C. 2001) analyse les problèmes de reconnaissance des termes coordonnés.

Exemples :
-Puissances de calcul et de stockage --> puissance de calcul et puissance de stockage.
-Elévateurs et abaisseurs de fréquence --> élévateurs de fréquence | abaisseurs de fréquence.

## III.7 Brève introduction à la terminologie informatique [12]

La terminologie informatique constitue l'ensemble des termes et des sigles utilisés dans le domaine de l'informatique. La terminologie informatique englobe des termes relatifs à des notions, des techniques, des normes, des produits – logiciels ou matériels, ainsi que des applications pratiques et des métiers de l'informatique. Elle est apparue dans les soixante dernières années, à l'origine de nombreux néologismes en anglais, qui ensuite a été adaptée dans d'autres langues, ce en utilisant théoriquement deux approches antithétiques :

Dans la langue italienne, roumaine ou encore russe, les termes en anglais on été repris tels quels. Dans ce cas on parle d'emprunt, qui est plus adapté.

Cependant, la langue française, pour créer son vocabulaire d'informatique a utilisé un volet divers de moyens d'enrichissement lexical qu'on retrouve probablement dans l'évolution lexicale de toutes les langues, mais a favorisé la traduction.
Puisque l'anglais de l'informatique a repris toute une série de termes de domaines contigus (la typographie, la presse, les mathématiques, etc.) le français, qui avait dans ces domaines une terminologie déjà bien fixée, s'est limité à proposer les termes équivalents, qui subissent ainsi, dans les deux langues, des extensions sémantiques.
Quand l'anglais a créé des termes spécifiques au domaine, ces mots ont été principalement traduits ou adaptés, faisant appel aussi à des moyens rhétoriques (comme les métaphores et les métonymies) pour plus d'expressivité et vivacité.

Au moment où la société a besoin, dans la communication, d'une nouvelle notion, la langue peut recourir :
- A l'emprunt d'un mot étranger, si la notion a été lexicalisée dans une autre langue ;
- A un développement à l'intérieur du système lexical propre par l'extension sémantique d'un lexème existant ou bien :
- Par l'invention d'un mot nouveau (cf. Humbley, 1974 ; Koch, 2002 ; Stanforth, 2002 ; Thibaut, 2004 ; Trotter, 2009 ; Iliescu et al. 2010).

## III.8 Termes empruntés d'autres langages sectoriels

La majorité des termes informatiques sont liés à la rédaction et mise en page des textes ainsi qu'aux fournitures de bureau. Le lexique français a simplement repris l'équivalent de ses propres termes, voyons cela avec les exemples ci-dessous :

-Typographie et publication de textes :

Exemples :

**angl.** *font* → **fr**. *police* ;
**angl.** *font size* → **fr**. *taille des polices* ;
**angl**. *bold* → **fr**. *gras* ;
**angl**. *italic* → **fr**. *italique* ;
**angl**. *underline* → **fr**. *soulignement* ;
**angl**. *parquet* → **fr**. *paquet* ;
**angl**. (*spell and grammar*) *checking* →**fr**. *vérification/ contrôle orthographique* ;
**angl**. *page layout* → **fr**. *mise en page* ;
**angl**. *header/ footer* → **fr**. *en-tête/ pied de page* ;
**angl**. *widows/ orphans* → **fr**. *veuves/ orphelines* ;
**angl**. *printer* → **fr**. *imprimante* ;
**ang**l. *copy* → **fr**. *copie* ;
**angl**. *copy – paste* → **fr**. *copier – coller* ;

-Mathématiques :

**angl.** *subscript/ superscript* → **fr.** *indice/ exposant* ;
**angl.** *algorithmics* → **fr.** *algoritmique*;
**angl.** *parallel* → **fr.** *parallèle* ;
**angl.** *matrix* → **fr**. *matrice*,
**angl.** *application* → **fr.** *application* ;

-Philosophie :

**angl.** *heuristics* → **fr.** *heuristique*

-Constructions :

**angl.** *bridge* → **fr.** *pont* ;
**angl.** *gateway* → **fr.** *passarelle* ;

-Economie :

**angl.** *management* → **fr.** *gérance*,
**angl.** *to export* → **fr.** *exporter*,
**angl.** *to import* → **fr.** *importer* ;

-Télévision :

**angl.** *scan* → **fr.** *balayage* ;

-Bureau :

**angl**. *label* → **fr.** *étiquette* ;
**angl.** *notepad* → **fr.** *ardoise* ;
**angl.** *organizer* → **fr.** *-agenda* ;
**angl.** *bookmark* → **fr.** *signet* ;
**angl.** *folder* → **fr.** *dossier* ;
**angl.** *file* → **fr.** *fichier*.

Dans ces exemples, des extensions sémantiques on été appliquées, basées en général sur des similitudes des fonctions remplies par les entités qui leur servent de référents. Prenons comme exemple, le suivant : « Quand quelqu'un interrompt la lecture d'un livre, il marque souvent l'endroit du volume où il s'est arrêté à l'aide d'un objet (un petit ruban ou une bande de papier ou de carton) appelé *bookmark* en anglais et *signet* en français. Ce petit objet lui assure un accès rapide à la page ainsi marquée quand la personne/ l'utilisateur reprend la lecture. Au moment où les informaticiens ont créé une commande permettant d'accéder vite à une section d'un document marquée auparavant par l'utilisateur, ils ont appelé ce moyen d'accès rapide *bookmark,* terme traduit naturellement en français avec *signet*. Les deux mots ont subi le même développement sémantique, basé sur une similitude fonctionnelle entre l'ancien référent (la bande de carton) et le nouveau référent (la commande informatique), ressemblance qui permet à l'usager de comprendre facilement le nouveau sens. Le trait permettant un accès rapide se trouve dans l'intersection sémique qui explique le passage métaphorique de la première à la deuxième signification, dans les termes du Groupe μ (1972). »[5]

## III.9 Intégration des termes anglais en français

L'intégration sémantique des termes anglais d'informatique en français a permis de discerner plusieurs procédés employés :

**III.9.1 Extension sémantique :** L'extension sémantique est utilisée lors de l'adaptation d'un mot existant dans les deux langues anglais et français, ce processus est rendu plus facile car environ 30% de mots anglais proviennent de l'ancien français. En effet, après plusieurs siècles de bilinguisme plusieurs lexèmes avec une même étymologie existent dans les deux langues. On a comme exemple les deux mots « menu » et « icône » :

**III.9.1.1 angl. *menu* fr. *menu :*** Dans l'étymologique, le mot *menu* dérive du latin *minutus* « petit, menu », qui est le participe passé du verbe *minuere* « diminuer ».

**FR : *menu*** (du lat. *minuere* « diminuer »>lat. *minutus* part. passé ; anc. fr. « sans importance » - *menu peuple, menu plaisirs*) 1. (*extension sémantique*$_1$) détail (*par le menu* « en détail ») ; 2. (*extension sémantique*$_2$) liste détaillée ; 3. (*extension*

[5] (Centre d'Études poétiques, Université de Liège, Belgique) poursuit depuis 1967 des travaux interdisciplinaires en rhétorique, en poétique, en sémiotique et en théorie de la communication linguistique ou visuelle, travaux qu'il signe d'un nom collectif (à l'instar de Bourbaki en mathématiques).

*sémantique*$_3$) liste détaillée des plats composant un repas (au restaurant) (DAF[6] 1718 cf. TLFi[7]) ; 4. (*extension sémantique*$_4$ *en informatique empruntée à l'anglais*) liste d'options offerte par un programme.

**ANGL : *menu*** (de l'anglo-normand) 1. sans importance ; 2. (*extension sémantique*$_1$ *empruntée au français*) liste des plats (1718 cf. *OED*[8]) ; 3. (*extension sémantique*$_2$ *en informatique*) liste d'options offerte par un programme.

Ce mot apparaît initialement en ancien français. L'adjectif connaît une première nominalisation, *le menu* signifiant « le détail », signification fréquente, puisqu'elle se trouve à la base de la locution adverbiale *par le menu* « en détail », locution attestée en 1538. Le sème « détaillé(e) » est le point de départ d'une spécialisation sémantique, le mot arrivant à désigner une liste détaillée et ensuite la « liste détaillée des plats composant un repas » (*Dictionnaire de l'Académie*, 1718, cf. *TLFi*). Le mot désigne maintenant surtout l'ensemble des repas servis dans un restaurant pour un prix déterminé, et, par métonymie, la carte sur laquelle figurent ces plats et leur prix. Une extension métonymique ultérieure a conduit à l'emploi de l'item pour nommer le repas même.

En anglais le mot était présent depuis longtemps, avec le sens « qui a peu d'importance, de valeur » (mot anglo-normand attesté en 1050, cf. *AND*[9], *OED*). Le syntagme *menu de repas* a été emprunté au français en 1837 (cf. *OnlineED*[10]). Dans les deux langues, le mot a connu un développement sémantique par généralisation, « liste détaillée de laquelle le client peut choisir un certain plat ». Au moment où *menu* a acquis le sens technique
de « liste d'options utilisées dans les programmes », les informaticiens français ont fait, tout naturellement, la même extension pour le mot déjà présent dans leur langue. Le *TLFi* lui donne comme synonyme le syntagme *liste d'options*.

**III.9.1.2 angl. *icon*, fr. *icône* :** Les mots *icon*, *icône*, sont un emprunt du grec byzantin et désignent une image sainte dans l'Église d'Orient.

**ANGL : *icon*** **(**du grec *eikon* « image, portrait ») 1. image sainte de l'Église d'Orient ; 2. (*extension sémantique en informatique*) image qui apparaît sur le moniteur et qui permet de donner une commande à l'ordinateur.
**FR : *icône*** **(**du grec *eikon* « image, portrait ») 1. image sainte de l'Église d'Orient ; 2. (*extension sémantique en informatique sous l'influence de l'anglais* – **emploi interdit**) image qui apparaît sur le moniteur et qui permet de donner une commande à l'ordinateur.

[6] DAF : Le dictionnaire de l'Académie Françoise
[7] TLFi : Trésor de la langue Française informatisé
[8] OED : Oxford English Dictionnary
[9] AND: The Anglo-Norman Dictionary
[10] OnlineED: Online Etymology Dictionary

Comme pour l'exemple précédent, les deux langues ont développé parallèlement un sens général, cette fois-ci celui de « image », ce qui a permis que le sens informatique du mot anglais (« graphique d'une commande sur le bureau ») soit repris en français aussi.

**III.9.2 Emprunt :** Comme on l'a déjà dit précédemment, les mots empruntés de l'anglais sont plus nombreux dans la langue, italienne, roumaine et russe ; Et sont réduits dans la langue française. Dans certaines situations le français a emprunté également des mots à l'anglais comme par exemple le mot « hacker » :

**III.9.2.1 angl. *hacker*, fr. *hacker* (*bidouilleur, fouineur*)**

C'est un mot anglais entré dans le vocabulaire informatique de beaucoup de langues. Il désigne une personne « qui visite et pirate les banques de données » et sa reprise a créé des difficultés et des hésitations aux informaticiens français, qui ont proposé plusieurs équivalents.

**ANGL : *hacker*** (de *to hack* + *-er* « taillader ») 1. (*développement argotique des étudiants de MIT)* détournement habile ; 2. (*extension sémantique en informatique*) délinquant informatique. **$FR_1$ : *hacker*** (*emprunt à l'anglais*) (*inform*.) délinquant informatique ;

**$FR_2$ : *bidouilleur*** (de *bidouiller* + *-eur*) 1. personne qui bricole, trafique ou triche ; 2. (*extension sémantique en informatique sous l'influence de l'anglais*) délinquant informatique ;
**$FR_3$ : *fouineur*** (nominalisation de *fouineur* « très curieux » de l'animal *fouine*) 1. (*extension sémantique informatique selon le modèle anglais*) délinquant informatique.

En anglais le substantif nom d'agent dérive du verbe *to hack* « taillader (une branche, un objet) en donnant des coups en biais ». Le mot acquiert le sens de « délit informatique » aux États-Unis en 1976 (cf. *OED*). Les extensions sémantiques du mot dans l'anglais étatsunien sont matière de débats. Selon une des hypothèses, il s'agit d'une signification donnée au mot dans l'argot estudiantin de MIT (Massachusetts Institute of Technology) dans les années 60. Dans cet argot, le mot signifie, en mathématiques, une solution ingénieuse à un problème, en droit, un détournement habile des prévisions d'une loi.

Beaucoup de sites Internet emploient le terme avec le sens de « copier, imiter, (extension) escroquer ». Le *Dictionnaire de l'informatique et de l'internet* (*DicoFr*) explique le terme anglais comme signifiant « bricoleur, bidouilleur » et désignant les programmateurs astucieux et débrouillards. Le sens actuel du mot en anglais est celui de « délinquant informatique », sens avec lequel le mot a été emprunté par de nombreuses langues.

En français on a essayé de donner deux traductions du substantif *hacker* : (i) *bidouilleur*, dérivé nominal du verbe *bidouiller*, qui a non seulement le sens de « bricoler » mais aussi celui de « trafiquer, truquer, tricher », et (ii) *fouineur*, nominalisation de l'adjectif *fouineur*, *-euse*, « personne très curieuse », du verbe *fouiner*, « se livrer à des recherches

méticuleuses », à son tour dérivé du substantif *fouine*, nom d'un petit mammifère carnivore qui est très curieux de nature (angl. *marten*). En revanche, le nom de l'action (*hacking*) a été traduit avec succès par *piratage* (métaphore dans un scénario ou un *frame*[3] de navigation, sur les mers et les océans ou bien sur l'Internet).

Pour la catégorie des emprunts, nous pouvons citer d'autres exemples : *pixel* (mot qui désigne dans les deux langues le plus petit élément d'une image caractérisé par la luminosité, la couleur, etc.) *to click → cliquer* (mot d'origine onomatopéique, signifiant l'emploi du bouton-poussoir d'une souris pour activer les diverses commandes), *scanner*

→ *scanneur* (appareil destiné au transfert de documents sur papier vers des fichiers numériques), *server → serveur* (système informatique qui permet la consultation des banques de données), etc.

**III.9.3 Traduction :** Un autre moyen employé pour la création du lexique français de l'informatique a été la traduction : le terme anglais (souvent une métaphore ou une métonymie d'un substantif concret) est traduit par le terme concret correspondant, le terme français acquérant lui aussi un sens figuré. Nous pouvons citer des mots comme :

**1-** *Mouse → souris*
(passage métaphorique dû à une analogie de forme et de mouvement) ;
***2****-Desktop → bureau* (métonymie basée sur une similitude des fonctions, le bureau traditionnel et le bureau informatique désignant le lieu où se trouvent les éléments nécessaires au travail – les dossiers, les fichiers, etc.) ;
**3-** H*ome page → page d'accueil* (métaphore pour la page de présentation d'un site) ;
*Gateway → passerelle* (métaphore désignant le dispositif qui permet la connexion à des réseaux) ;
**4-**I*nformation highway → autoroutes de l'information* (métaphore, pour les moyens d'informatique interconnectés) ;
***5-*** *Frame → cadre* (métaphore, pour désigner la sous-fenêtre qui affiche des documents) ;
***6-*** *Recycle bin → corbeille* (métaphore basée sur une similitude de fonctions) ;
***7-*** *Chat → causette* (communication par l'intermède des messages écrits sur les moniteurs entre plusieurs personnes sur l'Internet).

Ce dernier terme français a été abrogé par le *Journal Officiel* (5 avril 2006) qui a proposé le terme plus neutre, à « saveur » administrative, de *dialogue en ligne.* Pourtant, *causette*, étant un mot plus « sympathique », il continue à être utilisé, comme le prouve une recherche sur Google, qui le donne comme synonyme du verbe *chater* (une adaptation évidente de *to chat*).
Il s'agit en anglais d'un mot désignant un objet usuel qui, pour des raisons de forme et de fonction, subit une extension sémantique. En le traduisant par un mot désignant le même objet usuel en français, on reprend aussi le passage métaphorique, sémantiquement transparent :

**ANGL :** ***window*** (du moyen anglais *wyndouwe*) 1. ouverture couverte de vitres, faite dans un mur, une paroi, pour laisser pénétrer l'air et la lumière et permettant de voir dehors ;

l'encadrement de cette ouverture, souvent rectangulaire ; 2. (*extension sémantique métaphorique en informatique*) partie rectangulaire de l'écran d'un ordinateur à l'intérieur de laquelle sont affichées les informations relatives à une activité déterminée.

**FR : *fenêtre*** (du lat. *fenestra*) 1. ouverture couverte de vitres, faite dans un mur, une paroi, pour laisser pénétrer l'air et la lumière et permettant de voir dehors ; l'encadrement de cette ouverture ; 2. (*extension sémantique métaphorique en informatique sous l'influence de l'anglais*) partie rectangulaire de l'écran d'un ordinateur à l'intérieur de laquelle sont affichées les informations relatives à une activité déterminée.

**III.9.4 Invention :** Dans certains cas, au lieu de traduire, les informaticiens français ont inventé des mots nouveaux, qui présentent l'avantage de ne pas être ambigus, mais qui ont le désavantage de ne pas être transparents.

**III.9.4.1 angl. *e-mail*, fr. *courriel* :** C'est un des mots fréquents dans les deux langues, car ce moyen d'échange des messages est très employé, surtout par les générations d'âge moyen et dans beaucoup de milieux (affaires, universitaires, politiques, administratifs, etc.) :

**ANGL : *e-mail*** (abréviation de *electronic mail*, 1982) courrier électronique.

**FR : *courriel*** (mot-valise *courri*(*er*) + *él(ectronique)*, terme officiel) courrier électronique, *courrier/ message électronique*, *Mél* (« Le symbole : Mél., pour "messagerie électronique", peut figurer devant l'adresse électronique sur un document. "Mél" ne doit pas être employé comme substantif » (*Journal officiel* apud PR)).

Le mot *courriel* a été proposé par les informaticiens québécois comme équivalent français du mot anglais *e-mail*. Ce mot-valise est devenu officiel (le *Journal Officiel* du 20 juin 2003), mais il présente l'inconvénient de ne pas être compris par beaucoup de francophones, qui continuent à parler de *e-mail*.

**III.9.4.2 angl. *computer*, fr. *ordinateur* :** C'est, pour le français, un exemple des plus curieux, car il existe un contraste évident entre le « sentiment » linguistique des personnes qui réglementent la terminologie et la vraie situation linguistique du lexème.

**ANGL :** *computer* (du lat. *computere* « calculer » + *-er*) 1. personne qui faisait les calculs ; 2. (*extension sémantique*$_1$) mécanisme qui aide à faire des calculs (1869) ; 3. (*extension sémantique*$_2$ *en informatique*) machine électronique de traitement numérique de l'information (1944).

**FR :** *ordinateur* (du mot latin *ordo, ordinis* « ordre ») (en 1951 ou 1955) machine électronique de traitement numérique de l'information.

La sémantique du mot anglais a subi au début un passage métonymique habituel, de la personne qui fait un certain travail (dans ce cas des calculs, *OED*) à l'outillage qui fait la même opération (en 1869). Le sens actuel, de « machine électronique de traitement

numérique de l'information », apparaît pour la première fois en 1944, dans un rapport de John van Neumann, principal créateur du calculateur à programme.

Le mot français *ordinateur* est le nom donné par J. Perret aux premières machines de l'IBM (en 1955 selon *DicoFr*, et en 1951 selon le *PR*). C'est toujours le *Petit Robert* qui précise que le mot *ordinateur* (du latin *ordo, ordinis*) « remplace **l'anglicisme** *computer* ». Le nouveau terme s'est imposé, il existe même une abréviation familière, *ordi* (*brancher des ordis en réseau*, *PR*).

La partie comique dérive du fait que le lexème *computer* a été ressenti comme un anglicisme, raison qui a déterminé son élimination dans les années 50, au nom d'une bataille contre le franglais avant la lettre. On a ignoré le fait que le mot anglais dérive du verbe *to compute* qui existait déjà dans l'anglo-normand avec la signification

« calculer une date » et aussi « faire des additions ». Encore plus, le verbe *computer* avec le sens de « déterminer, calculer une date » existait déjà chez Montaigne (1595), comme emprunt savant du latin *computare*. En éliminant *computer*.

**III.9.5 Calques sémantiques, métaphores et adaptations :** Parfois le passage des termes d'informatique de l'anglais en français implique des processus sémantiques complexes, montrant la créativité (non seulement technique mais aussi lexicale) et parfois le sens de l'humour de leurs auteurs. Comme les exemples cités ci-dessous :

**III.9.5.1 angl. *phishing*, fr. *hameçonnage, filoutage :*** Une des fraudes informationnelles les plus fréquentes sur l'Internet consiste dans la tentative d'apprendre, auprès d'internautes crédules, des informations confidentielles, telles que des mots de passe pour les services en ligne, les numéros des cartes bancaires, etc. Cet acte de piraterie est désigné par le mot anglais *phishing*, en français par les mots *hameçonnage* ou *filoutage*.

**ANGL : *phishing* (**contamination de *phreak* « parler au téléphone sans payer » + *fishing* « l'action de pêcher » 1996) fraude sur l'Internet, spécialement en se faisant passer pour une compagnie réputée pour persuader un individu à révéler des informations secrètes, comme ses mots de passe, le numéro de ses cartes de crédit, etc.

**$FR_1$ : *hameçonnage*** (de *hameçon* « petit crochet de métal attaché à une ligne auquel on fixe l'appât ») 1. (*extension sémantique*) appât: *avaler l'hameçon, mordre à l'hameçon* ; 2. fraude sur l'Internet, spécialement en se faisant passer pour une compagnie réputée pour persuader un individu à révéler des informations secrètes, comme ses mots de passe, le numéro de ses cartes de crédit, etc.
**$FR_2$ : *filoutage*** (du verbe *filouter* « escroquer, voler par ruse, par tromperie ») fraude sur l'Internet, spécialement en se faisant passer pour une compagnie réputée pour persuader un individu à révéler des informations secrètes, comme ses mots de passe, le numéro de ses cartes de crédit, etc.

Le mot anglais *phishing* semble résulter de plusieurs « contaminations » : on part de *phreak* « parler au téléphone sans payer » et de son homophone, le substantif *fishing*, qui

désigne l'action de pêcher (cf. *OED*). Le mot présente, donc, deux éléments sémantiques fondamentaux : l'idée de fraude (bénéficier d'un service sans payer son coût) et l'idée de tromper par des ruses une future victime. Seulement un des deux éléments se retrouve dans les traductions proposées en français.

Une première traduction considère fondamental un élément de la *qualia structure* (Pustejovsky, 1995)[11] du mot *fishing*, à savoir l'appât utilisé par le pêcheur pour prendre sa proie. Ce fait explique la traduction *hameçonnage*. En plus, par un développement métonymique (de soutient à l'objet soutenu) *hameçon* signifie aussi la mouche ou le ver qui attire le poisson et, par un passage métaphorique, « apparence trompeuse, artifice destiné à attirer et à séduire quelqu'un » ; le mot est, donc, synonyme de *appât* (*TLFi*).

Ce sens se retrouve dans les expressions figurées *avaler l'hameçon, mordre à l'hameçon*, correspondantes de l'expression anglaise *to take the bait*. Une autre traduction française est *filoutage*, du verbe *filouter*, « escroquer, voler par ruse, par tromperie » (*TLFi*). Donc, du point de vue sémantique, les mots *phishing* et *hameçonnage*, respectivement *filoutage*, ont en commun seulement un trait très général : « une action délictueuse ». Le mot anglais contient aussi l'idée de « connexion gratuite », que les traducteurs français ont négligé, insistant sur l'autre aspect, celui de la manipulation, de la tentative de circonvenir la possible victime (dans un scénario de pêche ou non).

## III.10 Termes d'informatiques en linguistique

Le PC étant devenu un outil quasi universel pour toutes les personnes qui communiquent par écrit les résultats de leurs recherches, il est, en quelque sorte, normal qu'une partie de ses termes spécifiques passent dans d'autres domaines. Ici on a comme exemple, deux termes d'informatique entrés dans la terminologie linguistique actuelle.

**III.10.1 angl. *by default,* fr. *par défaut :*** Le syntagme du langage des informaticiens *by default* a été repris par la linguistique actuelle, pour désigner des valeurs sémantiques et pragmatiques fondamentales, qui peuvent être modifiées par le contexte.

**ANGL : *(by) default*** (moyen anglais < ancien français *defaute* dérivé de *defaillir* dérivé à son tour de *faute* et *faillir* ; les formes *defalte, defaulte* sont attestées dans l'anglo-normand au XIII$^{e}$ siècle) 1. carence, manque, pénurie ; 2. imperfection, défaut, tare ; 3. panne, échec ; *by default* (*droit*) échec dans l'accomplissement d'obligations légales ; (*inform*.) option sélectée d'avance par l'ordinateur si l'utilisateur ne choisit pas une option alternative. (*OED*) ; (*ling*.) valeur sémantique et/ ou pragmatique fondamentale d'un élément linguistique qui peut être modifiée par le contexte.

**FR : *(par) défaut*** (de l'ancien participe passé de *défaillir* au sens de « manquer », milieu du XII$^{e}$ s.) 1. absence de ce qui serait nécessaire ou désirable ; 2. (*droit*) situation d'une partie qui s'abstient d'accomplir les actes de la procédure dans les délais requis ou ne se présente pas à l'audience ; (*PR*) ; (*ling*.) valeur sémantique et/ ou pragmatique fondamentale d'un certain élément linguistique qui peut être modifiée par le contexte.

[11] Pustejovsky, J. (1995) The Generative Lexicon, MIT Press, Cambridge, MA.

Le mot anglais *default*, comme *menu*, provient de l'ancien français, étant présent dans l'anglo-normand, dans les *Gloss Nequam* (environ 1200) avec le sens de « carence, pénurie ; pêché » (*AND*). En français le mot dérive de l'ancien participe passé de *défaillir* au sens de « manquer » (fin XI$^{e}$ siècle, cf. *PR*). À son tour *défaillir* provient du latin populaire *fallire* « tromper, échapper à ». Le *TLFi* signale en français l'emploi du mot non seulement dans la langue courante, mais aussi dans plusieurs domaines techniques : droit (*jugement par contumace*/ *par défaut*, emploi qui se retrouve en anglais aussi, *judgement by default*), chasse (pour désigner la perte de la piste de l'animal pourchassé), en psychologie (*défaut d'expérience*, *de mémoire*), en mathématiques et en physique nucléaire (où le mot signifie « différence »). En anglais, l'éventail des sens semble être plus restreint, étant dominé par la signification de « manque », de « défaillance » (dans l'action, dans les payements), ou de « absence » (de l'équipe adverse, d'un candidat valable). En informatique, les expressions *by default*, respectivement *par défaut*, désignent la valeur attribuée automatiquement par un logiciel en l'absence d'une indication différente et explicite de l'utilisateur (*mise en page par défaut*, *police de caractère par défaut*, *langue par défaut*, *clavier par défaut*, etc.).

En linguistique, le terme *par défaut* désigne la valeur sémantique fondamentale d'un certain mot ou d'une certaine catégorie grammaticale, valeur qui peut être changée par le contexte. Par exemple, la valeur *par défaut* du présent de l'indicatif est celle d'exprimer la simultanéité avec le temps d'énonciation, donc la simultanéité déictique.

Cette valeur peut être modifiée par le contexte : il existe un présent omnitemporel (*deux et deux font quatre*), un présent à valeur de futur (*j'arrive dans un quart d'heure*), le présent narratif, souvent sans valeur temporelle, comme dans les histoires drôles (*John entre dans un pub*), le présent d'une indication scénique (*Marie époussette*), le présent d'une prévision (*le train part à deux heures trente-huit*), le présent d'une prédication itérative (*chaque soir, il lit quelques pages en russe*), etc. Le terme *par défaut* permet aussi de faire la différence entre la signification conceptuelle (*par défaut*) en opposition avec la signification instructionnelle des divers éléments linguistiques.

**III.10.2 angl. *interface*, fr. *interface* :** Le mot *interface* a été créé pour la première fois en anglais (attesté en 1882, cf. *OED*) pour désigner une surface qui se trouve à la frontière entre deux parties de matière ou d'espace.

**ANGL** : (de *inter* préfixe + *face* nom, 1882) 1. surface qui se trouve à la frontière entre deux parties de matière ou d'espace ; 2. moyen ou lieu d'interaction entre deux systèmes, organisations, etc. ; lieu de contact entre deux systèmes ou disciplines ; interaction, lien, dialogue (*OED*) ; 3. (*inform.*) une voie d'échange d'informations qui permet de communiquer à un ou plusieurs ordinateurs ou à des équipements extérieurs (imprimantes, moniteurs, modems) (*DicoFR*) ; 4. (*ling.*) valeur sémantique et pragmatique fondamentale d'un élément linguistique qui peut être modifiée par le contexte.

**FR** : (empr. à l'angl. *interface* « surface à la frontière entre deux parties de matière ou d'espace » (1882) d'où « espace, lieu d'interaction, moyens d'interaction, de jonction entre deux systèmes, deux organisations » 1962). 1. (*chim.*) surface de contact entre deux

milieux ; 2. (*informat*.) jonctions entre deux logiciels leur permettant d'échanger des informations ; (*au fig*.) zone de contact et d'échanges ; 3. (*ling*.) valeur sémantique et pragmatique fondamentale d'un élément linguistique qui peut être modifiée par le contexte.

En anglais, le mot composé est formé du préfixe d'origine française *inter-* et du mot *face*, qui provient d'un mot anglo-normand signifiant « visage » (attesté au XII$^{e}$ siècle). Le mot *interface* est repris en français, comme emprunt de l'anglais, en 1965 (cf. *TLFi*).

En informatique le terme signifie, dans les deux langues, la partie visible qui permet à l'utilisateur d'un logiciel de gérer ses interactions avec la machine, ainsi que les équipements qui permettent à deux ou plusieurs ordinateurs ou à leurs équipements extérieurs de communiquer (*DicoFR*). Par extension, le terme a acquis la signification plus générale de « zone de contacts et d'échanges » (*TLFi*).

En linguistique, le terme a été proposé pour la première fois par Nicholas Asher et Alex Lascarides au début des années 90, sous la forme de *l'interface pragmatique – sémantique*, théorie qui représente un développement de la sémantique dynamique de Hans Kamp. L'emploi du terme est en expansion ; par exemple dans l'encyclopédie de pragmatique éditée par Laurence Horn et Gregory Ward, il y a toute une section, « Pragmatics and its Interfaces » (Horn & Ward, 2004 : 405–604), où divers auteurs débattent les rapports entre la pragmatique et d'autres domaines (grammaire, sémantique, lexique, l'intonation, l'apprentissage de la langue, la philosophie du langage, la linguistique computationnelle).

Comme, pour reprendre la célèbre phrase de Ferdinand de Saussure, « la langue est un système où tout se tient », les linguistes n'ont pas attendu l'apparition de l'informatique pour étudier les rapports qui existent entre les divers niveaux du langage (phonologique, morphologique, syntaxique, etc.). Déjà l'école distributionaliste américaine a montré qu'il n'existe pas de frontière entre la morphologie et la syntaxe. La discipline qui a résulté de cette constatation a été appelée « morphosyntaxe ». Il existe aussi une phono-morphologie. La procédure de réduire le nom d'une des disciplines en préfixe a été employée aussi pour désigner les zones communes de deux disciplines différentes (comme « sociolinguistique », « psycholinguistique »). Une autre procédure de création terminologique pour ces zones linguistiques « de contact » est celui de transformer le nom d'un des domaines en adjectif (« sémantique lexicale ») où en complément du nom (« philosophie du langage »).

À présent, sous l'influence de l'informatique, ces disciplines s'appelleraient « interface morphologie - syntaxe » ou bien « interface phonologie - morphologie », « interface linguistique - psychologie », etc. Du point de vue lexical, la création d'un syntagme au lieu d'un substantif composé présente l'avantage de la flexibilité : il serait difficile de parler d'une « sémantosyntaxe » (qui s'appelle, en fait, « sémantique combinatoire ») ou d'une « pragmatomorphologie », tandis que les syntagmes avec le substantif « interface » nous permettent de créer un grand nombre de termes qui ne sont pas rébarbatifs et qui sont, en même temps, transparents.

## III.11 Conclusion

Comme on l'a vu dans ce troisième chapitre, le langage de l'informatique a été créé en anglais par la reprise, l'extension sémantique, la métaphorisation des termes déjà existants dans la langue ou bien par l'invention de termes nouveaux. Les mêmes méthodes ont été employées aussi en français, auxquelles on peut ajouter l'emprunt à l'anglais, qui est assez limité. En effet ce vocabulaire sectoriel du français a été traduit et adapté de l'anglais.

Le problème qui se pose est celui de choisir entre la reprise telle quelle du mot étranger, sa traduction, la création d'un mot nouveau ou l'élaboration d'un passage métaphorique ou métonymique similaire. Chaque langue semble recourir, en proportions différentes, à tous ces moyens.

Par conséquent la diffusion du PC et de son emploi par des chercheurs de, pratiquement, tous les domaines de la science, les termes d'informatique commencent à être repris dans d'autres sciences, enrichissant ainsi la terminologie scientifique et technique d'autres domaines.

# *Chapitre IV*

## *Le système NooJ*

# Chapitre IV: Le système NooJ [10]

## IV.1 Introduction

Ce quatrième chapitre fait l'objet d'une présentation complète et détaillée du système NooJ. Cet outil est utilisé dans des applications variées du Traitement Automatique des Langues Naturelles (par ex. moteurs de recherche, extracteurs d'entités nommées, terminologie, traduction automatique), ainsi comme environnement de développement linguistique permettant de construire et de gérer des dictionnaires et grammaires électroniques, afin de formaliser divers niveaux des langues naturelles : orthographe, morphologie flexionnelle et dérivationnelle, lexique de mots simples, mots composés et expressions figées, syntaxe locale et désambiguïsation, syntaxe structurelle et transformationnelle, sémantique et ontologies.

## IV.2 NooJ : une plateforme de développement linguistique

NooJ est un environnement linguistique de développement (Silberztein, M.2003)[12]. Les fonctionnalités de NooJ sont adaptées à un public très varié incluant des linguistes (description de la morphologie et de la syntaxe des langues, analyse de corpus), des informaticiens du TAL (applications d'extraction d'information), ainsi que des enseignants tant en Linguistique qu'en enseignement des langues. Par exemple l'utilisation de NooJ pour l'enseignement des langues, plus particulièrement l'apprentissage du français langue étrangère (Silberztein M. et Tutin A. 2005).

NooJ est utilisé pour la construction à grande couverture de descriptions formalisées des langues naturelles dans le but de les appliquer en temps réel à de gros corpus. Les descriptions des langues naturelles sont formalisées dans des dictionnaires électroniques par des grammaires représentées par des ensembles de graphes.
NooJ fournit des outils pour décrire :

1- La morphologie flexionnelle et dérivationnelle,
2- Les variations terminologiques et orthographiques,
3- Le vocabulaire (mots simples, mots composés, expressions figées),
4- La syntaxe
5- et la sémantique.

**IV.2.1 Architecture intégrée :** Dans NooJ, toutes les connaissances linguistiques sont séparées de l'algorithme d'analyse lui-même. L'architecture globale de NooJ est basée sur un ensemble de modules linguistiques : orthographique, flexionnel, morphologique, dérivationnel et syntactico-sémantique. Un ensemble d'outils est fourni :

1- Un éditeur de graphes,
2- Un concordancier,
3- Un débogueur de grammaires,
4- Des outils statistiques,

[12] Silberztein M. 2003. NOOJ'manual. http://www.nooj4nlp.net

5- Etc.

Cet ensemble en interaction permet de répondre aux besoins les plus pointus des utilisateurs de NooJ et notamment des spécialistes du traitement des langues (TAL).

**IV.2.2 Architecture orientée objet :** NooJ est développé sous une architecture orientée objets (Silberztein M. 2004). Les composants du système (objets) intègrent les données ainsi que les routines nécessaires pour leurs traitements. La communication et la coordination entre les objets sont réalisées par un mécanisme interne. Cette architecture présente les avantages suivants :

1- Eviter la redondance dans le code source grâce au concept d'héritage. Ceci rend le code source plus lisible et gérable ;
2- Accéder directement à toutes les méthodes publiques dans NooJ, au lieu d'avoir à sélectionner un ensemble de sous-programmes. Ceci apporte une flexibilité et une utilisabilité supplémentaires au système ;
3- Pouvoir ajouter d'autres fonctionnalités supplémentaires ou des méthodes spécifiques à des langues sans avoir à modifier l'architecture globale de l'application.

**IV.2.3 Utilisation de la technologie à états finis :** NooJ contient des outils permettant d'éditer, de tester, de déboguer et de gérer des ensembles importants de graphes à états finis. Tous les objets traités par NooJ sont ou peuvent être transformés sous forme de transducteurs à nombre fini d'états. Toutes les opérations sur les textes, grammaires et dictionnaires se ramènent ainsi à des opérations sur des transducteurs. NooJ utilise aussi des outils plus puissants comme les graphes récursifs qui sont équivalents à des grammaires hors-contexte.
NooJ ne traite pas que des graphes d'états finis, avec NooJ on peut aussi créer :
- Des grammaires algébriques ;
- Des graphes récursifs : RTN (Recursive Transition Network).

**IV.2.4 Développement de ressources linguistiques à large couverture :** Les différentes ressources linguistiques sont décrites dans des structures de données autonomes (Silberztein, 2005). En l'occurrence, elles sont représentées soit dans des formats textuels lisibles et accessibles par un non-informaticien (comme pour les dictionnaires et les fichiers des paradigmes flexionnels et dérivationnels), soit sous forme de graphes facilement accessibles et compréhensibles qui peuvent aisément être contrôlés et modifiés sans avoir à maîtriser l'ensemble du programme (comme pour les grammaires flexionnelles, morphologiques et syntaxiques).

**IV.2.5 Traitement de corpus :** NooJ peut traiter directement des ensembles potentiellement importants de documents. Ces documents peuvent être codés dans plus d'une centaine de formats :
- tous les formats des types ASCII, EBCDIC, Unicode (UTF-8, UTF-16, etc.),

- les formats standards tels que HTML (format par défaut des textes disponibles en ligne), XML (format de bases de données textuelles), PDF (format de documents portables), RTF (format de textes enrichis), Microsoft WORD, etc.

**IV.2.6 Construction, édition et gestion de concordances sophistiquées :** NooJ permet de construire et de gérer des concordances sophistiquées. Il permet de combiner plusieurs requêtes, filtrer manuellement les résultats incorrects et de sauvegarder la concordance résultante. Les outils proposés pour la gestion des concordances permettent d'aborder certains problèmes linguistiques de façon incrémentale.

Par exemple, il est envisageable de commencer son travail avec NooJ par l'identification de certains termes ou expressions dans un corpus, tout d'abord de façon naïve (par exemple, rechercher tous les mots finissant par –ment pour identifier des adverbes en français, ou rechercher tous les nombres pour identifier des dates). Puis, au vu de la concordance résultante, il est possible de raffiner progressivement la caractérisation des phénomènes, d'une part en améliorant les requêtes (par ex. en tenant compte du contexte des mots), d'autre part en classifiant les résultats obtenus (par ex. pour distinguer les adverbes en –ment, par exemple "régulièrement", des noms en "–ment", par ex. "investissement"). Les résultats des requêtes peuvent être facilement contrôlés.

**IV.2.7 Annotation interactive de corpus :** NooJ permet d'appliquer des grammaires représentées sous forme de graphes ou d'expressions rationnelles augmentées à un corpus. Le résultat est affiché dans une table de concordances où la colonne du milieu contient le mot ou la séquence de mots reconnus. L'utilisateur peut valider ou éliminer chaque entrée de cette concordance. La concordance ainsi filtrée peut, ensuite, être utilisée pour annoter le corpus.

A ce propos, l'annotation d'un corpus à partir d'une concordance permet aux utilisateurs de vérifier, adapter et corriger les résultats d'un traitement automatique dont les sorties ne sont, pratiquement, jamais parfaites. Ceci est très important surtout lors d'une étape de préparation d'un corpus.

## IV.3 Les dictionnaires NooJ

Les dictionnaires NooJ (Silberztein M. 2003) contiennent indistinctement des mots simples et des mots composés qui peuvent représenter des variantes terminologiques ou phonétiques des entrées lexicales.

**IV.3.1 Les ALUs (Atomic Linguistic Units) :** Les unités atomiques linguistiques (ALUs) sont les plus petits éléments qui forment une phrase. Ce sont des "mots" dont la signification ne peut pas être calculée ou prédite, on doit les apprendre pour être capables de les utiliser. On doit décrire explicitement toutes les propriétés syntaxiques et sémantiques de ces mots afin de pouvoir les analyser.

D'un point de vue formel, NooJ sépare les ALU en quatre classes :

1- Les mots simples, ALU orthographiés sous forme de mots.

2- Les affixes, qui sont des ALU orthographiés comme des séquences de lettres, habituellement à l'intérieur des formes de mots, telles que dans "dé-" et "-isation" dans "dé-central-isation" ;
3- Des mots composés, ALU orthographiés par des séquences de formes de mots, tels que "table ronde", "pomme de terre" ;
4- Des expressions figées.

Afin de reconnaitre les mots simples et les affixes, Le système NooJ effectue une recherche dans les dictionnaires. Les autres types seront identifiés en découpant les formes de mots en plus petites unités. Les quatre formes sont reconnues par NOOJ en appliquant des grammaires syntaxiques.

### IV.3.2 Ressources pour reconnaître les unités linguistiques atomiques

En ayant choisi la langue française (fr), comme suit:

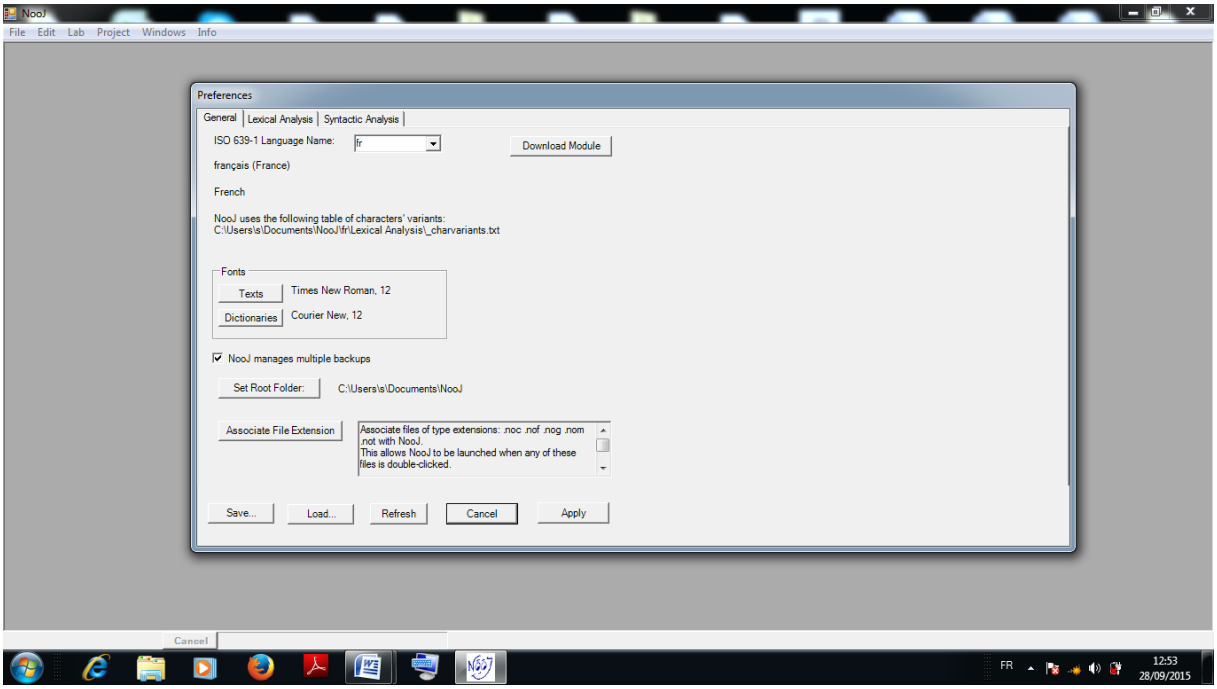


**Figure IV.3.2 (a) : "Choix de la langue".**

Dans NooJ, la commande :
Info > Preferences > Lexical Analysis, permet d'afficher les zones : "Dictionary" comme on peut le voir sur la figure suivante :

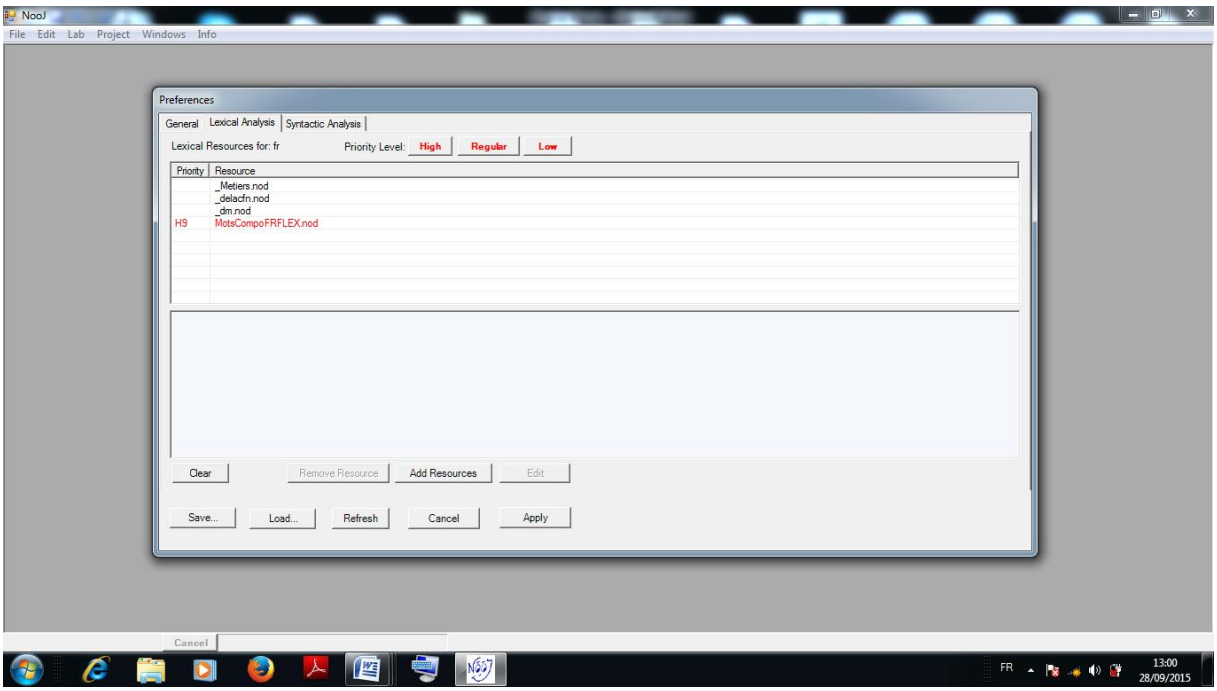


**Figure IV.3.2 (b) : Ressources lexicales: "Dictionary"**

La zone "Resource" contient toutes les ressources lexicales qui sont utilisées pour reconnaître les mots simples et les mots composés.

**IV.3.2.1 Les Dictionnaires :** Les dictionnaires NooJ sont des fichiers ".nod" qui sont compilés à partir de fichiers source ".dic". Techniquement les fichiers ".nod" sont des transducteurs d'états finis, ils proviennent de fichiers de type texte ".dic" qui sont compilés en utilisant la commande

Lab > Dic > compile

On utilise la commande Lab > Dictionary **(Figure IV.3.2.1 (a))** pour :

1- Fusionner les informations de deux dictionnaires ("Enrich"),
2- Produire automatiquement la liste de toutes les formes fléchies et dérivées correspondant aux entrées lexicales d'un dictionnaire ("Inflect"),
3- Compiler le dictionnaire en un fichier binaire .nod, ainsi il peut être appliqué à des textes, ("Compile").

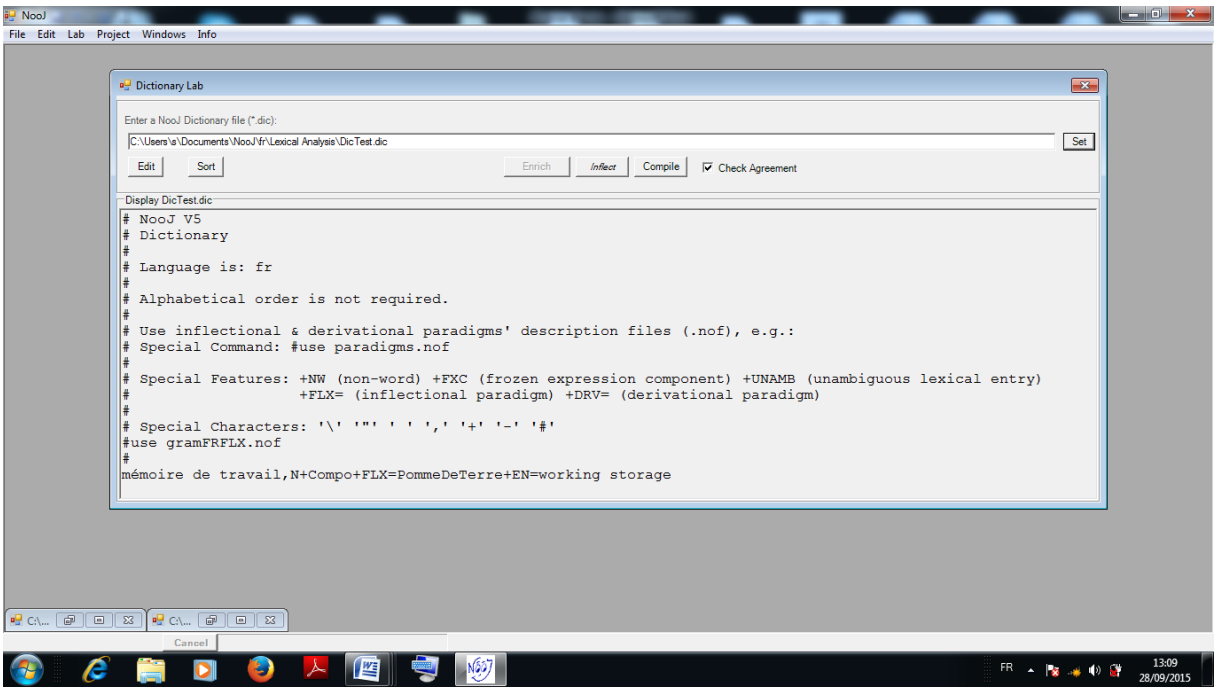


**Figure IV.3.2.1 (a) : Utilisation de Lab > Dictionary et "DicTest.dic".**

Dans la figure suivante, nous avons un exemple de dictionnaire, avec quelques entrées :

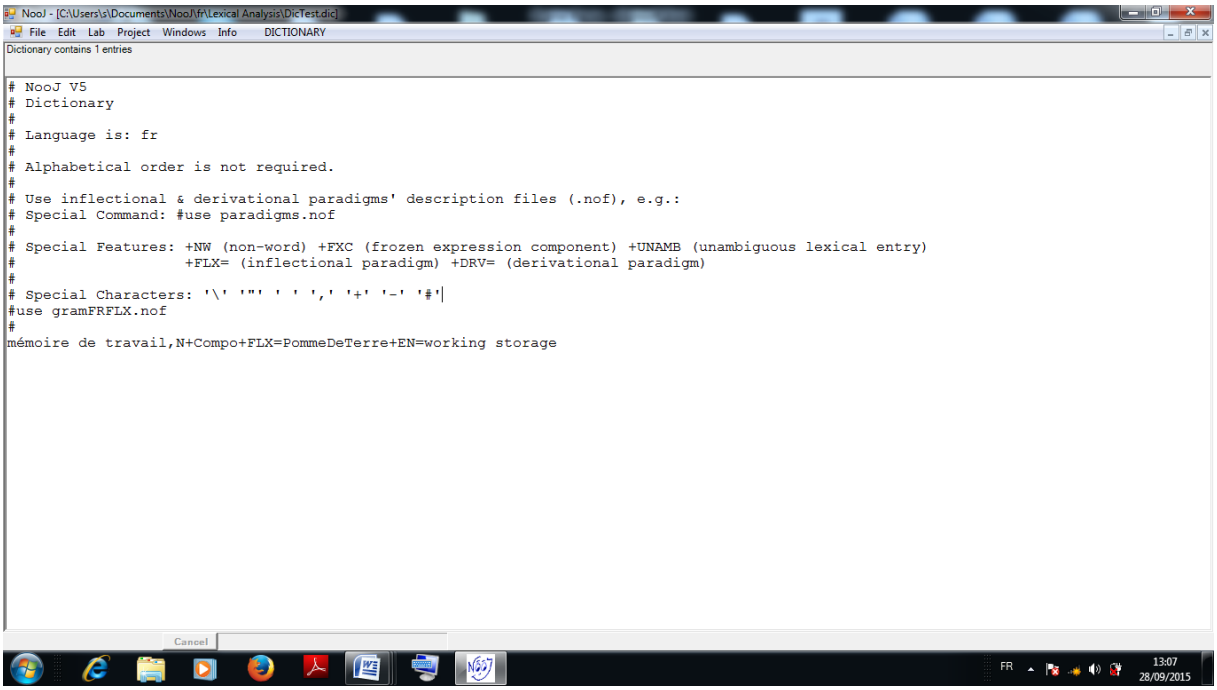


**Figure IV.3.2.1 (b) : Exemple de dictionnaire "DicTest.dic".**

**IV.3.2.2 La Morphologie :** La zone "Morphology" affiche toutes les grammaires morphologiques qui sont utilisées pour reconnaître les formes de mots à partir de leurs composants (préfixes, affixes et suffixes). Ces ressources morphologiques sont des fichiers ".nom" constitués d'ensembles structurés de graphes qui décrivent des paradigmes morphologiques.

**IV.3.3 Outils pour décrire la morphologie :** NooJ (Silberztein M. 2003) offre deux outils équivalents pour représenter et décrire la morphologie flexionnelle et dérivationnelle. Ces deux descriptions de la morphologie sont lexicalisées : chaque entrée dans un dictionnaire NooJ est associée avec un ou plusieurs paradigmes flexionnels ou dérivationnels. Les paradigmes sont décrits et stockés dans des grammaires flexionnelles et dérivationnelles sous forme de fichiers ".nof". NooJ offre aussi des grammaires morphologiques (fichiers ".nom).

**IV.3.3.1 Descriptions flexionnelles et dérivationnelles :** Ce sont des fichiers ".nof" organisés en règles qui décrivent les propriétés de chaque catégorie.

Exemple de règle :

pomm = <E>/f+s + s<P>/f+p;

"<E>/f+s" indique qu'on ne rajoute rien "<E>" au féminin "f" singulier "s", nous obtenons pour le mot "pomme" :

" s<P>/f+p ", indique <P> : aller à la fin du mot précédent, rajouter s, nous obtenons

+f+s => "pommes" au féminin pluriel.

**IV.3.3.2 Grammaires flexionnelles et dérivationnelles :** Elles sont représentées par des ensembles structurés de graphes qui décrivent des paradigmes morphologiques stockés dans des fichiers ".nof".

## IV.4 Format des dictionnaires NooJ

Généralement, le dictionnaire d'un langage donné contient tous les lemmes du langage et les associe avec un code morpho-syntaxique, des codes sémantiques et syntaxiques possibles, des paradigmes flexionnels et dérivationnels.

**IV.4.1 Exemples d'entrée :** Voici un exemple d'entrée :

pomme,N+FLX=pomm+Com

- L' entrée " pomme " est un nom, (catégorie "N"), elle se fléchit :

("pomme, pommes") selon le modèle "pomm" qui est la valeur de la propriété "FLX", le trait sémantique est Commestible ("Com").

**IV.4.2 Informations linguistiques :** Dans les dictionnaires NooJ, toutes les informations linguistiques telles que les codes syntaxiques tels que "+tr" (transitif) et les codes sémantiques tels que "+conc" (concret) et "+abst" (abstrait) doivent être précédés du caractère "+".

**IV.4.3 Codes d'information spéciaux :** Le code "+ NW " (Non-Word) est utilisé pour décrire des entrées lexicales abstraites qui n'apparaissent pas dans les textes et qui ne devraient pas être analysées comme des formes réelles de mots. Cette forme est utile pour la construction d'un dictionnaire dans lequel des entrées sont des stems ou des racines de mots.

Le code "+UNAMB" (unambigous ou mot non ambigu) indique à NooJ de stopper l'analyse de la forme du mot avec soit les ressources lexicales soit l'analyseur morphologique.

Le code "+FLX" (Paradigme flexionnel) permet d'indiquer le nom des règles de flexions du mot. Par exemple dans l'entrée de dictionnaire :

pomme,N+FLX=pomm+Com

pomme est associé avec la propriété +FLX=pomm, et pomm est un paradigme flexionnel.

Le code "+DRV" (Paradigme dérivationnel) permet aux lexicographes d'associer chaque mot avec les variantes dérivationnelles correspondantes. Des dérivations morphologiques seront ensuite décrites avec précision pour chaque entrée.
Lier des formes dérivées dans le dictionnaire permettra d'avoir un dictionnaire moins volumineux.

**IV.4.4 Propriétés lexicales :** Non seulement des formes, mais aussi des propriétés peuvent être associées aux entrées lexicales. Une propriété lexicale a un nom et une valeur exprimés sous la forme " +nom=valeur".

**IV.4.5 Variantes lexicales :** Les dictionnaires NooJ ne sont plus limités à un seul champ : ils peuvent contenir des entrées liées à un «super-lemme», qui peut être une variante orthographique, la traduction dans une autre langue, une entrée synonyme ou hyperonyme.

## IV.5 Fichiers de définition des propriétés ".DEF"

Dans le dictionnaire, l'instruction suivante indique où sont définies les associations catégories, propriétés et leurs valeurs :

#use _Exemple.def

Les utilisateurs peuvent créer de nouveaux codes d'information tels que "+Info" "+Politiqu" et peuvent les ajouter a des entrée lexicales. On peut aussi utiliser ces codes

d‘informations qui sont ajoutés sous forme de propriété avec une valeur, telle que "+Domain=Info".

Exemples :
- V_Nb = s|p signifie que le nombre du verbe est singulier ou pluriel ;
- V_Pers = 1|2|3 signifie que la personne du verbe est 1ère, 2ème ou 3ème.

## IV.6 Représentation formelle des dictionnaires électroniques

Les automates d’état finis (FSA : "finite state automata") sont utilisés dans beaucoup d‘applications de traitement du langage naturel (NLP). En particulier, ils fournissent un stockage et une récupération efficaces des ensembles finis de chaînes sur un alphabet fini, et peuvent donc être utilisés pour représenter le vocabulaire d'une langue.

L'utilisation des automates peut être étendue à l'utilisation des transducteurs d‘état finis (FST : "finite state transducer") qui permet aux utilisateurs d'associer à chaque élément d'un vocabulaire des informations, telles que des informations morpho-syntaxiques (genre, nombre, etc.), des informations syntaxiques et sémantiques (par exemple, transitif, humain, etc.), ou plus complexes, telles que la traduction des termes dans d'autres langues, un ensemble d'expressions synonymes, etc.

Le but d'une analyse lexicale d'un texte est de mettre en correspondance les termes de ce texte avec ceux d'un vocabulaire généralement décrits dans un dictionnaire. Dans la plupart des langues, les tokens sont des formes de mots fléchis et / ou dérivés qui nécessitent d‘être associés au lexème correspondant (entrée de dictionnaire) avec quelques informations linguistiques.

Ainsi, les transducteurs sont bien adaptés pour l‘exécution d‘analyses automatiques lexicales. Les unités linguistiques de NooJ sont reconnues en effectuant une recherche dans les dictionnaires, et aussi bien en utilisant des grammaires morphologiques pour analyser les formes du mot, ainsi qu‘en utilisant les grammaires syntaxiques des phrases à analyser.

Sur le plan linguistique, l'intérêt des graphes ou des automates est de regrouper dans une même description des éléments équivalents du point de vue du sens, mais différents au niveau de la forme, par exemple les variantes graphiques de mots dont l'orthographe n'est pas normalisée ou bien des variantes d'expressions de temps.

## IV.7 Exemples d'utilisation

**IV.7.1 Analyse lexicale :** On effectue l'analyse lexicale du texte "fruits.not", dans la **Figure IV.7.1**

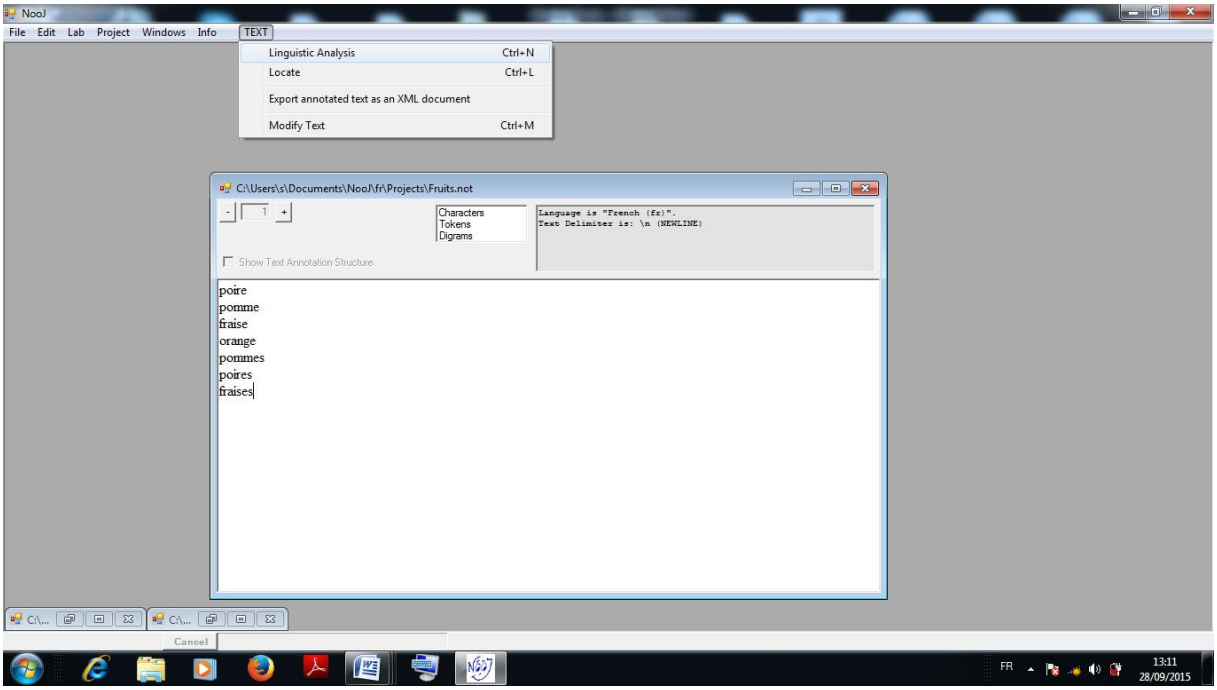


**Figure IV.7.1 : analyse lexicale du texte "Fruits.not"**

**IV.7.2 Affichage des annotations après analyse lexicale :** Dans la **Figure IV.7.2**, après l'analyse lexicale du texte "fruits.not", on affiche les annotations qui sont rangées dans un dictionnaire.

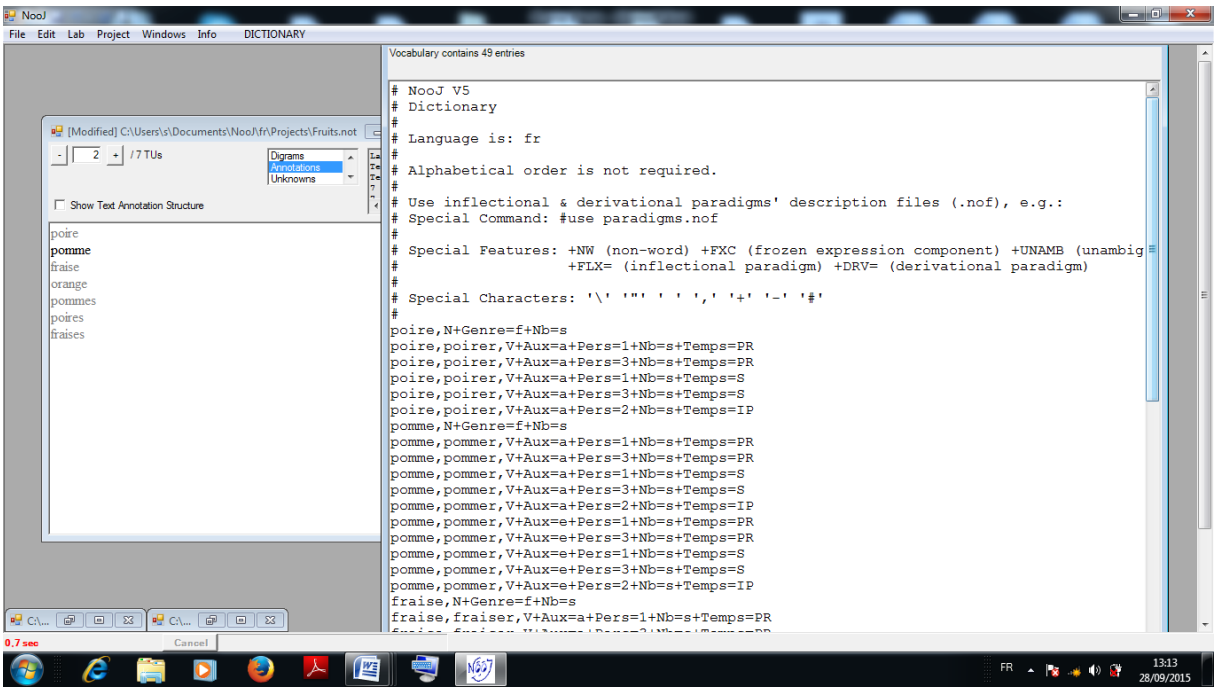


**Figure IV.7.2 : Annotations**

**IV.7.3 Recherche d'un "patron" :** Dans la **Figure IV.7.3**, une expression régulière est utilisée pour rechercher un patron dans le texte, ici <N+f+s> est une expression régulière NooJ utilisée pour trouver tous les noms (N), au féminin (f) singulier (s) dans le texte "fruits.not". La concordance est affichée et peut être sauvée dans un fichier, cela peut être utile pour l'acquisition de nouveaux mots.

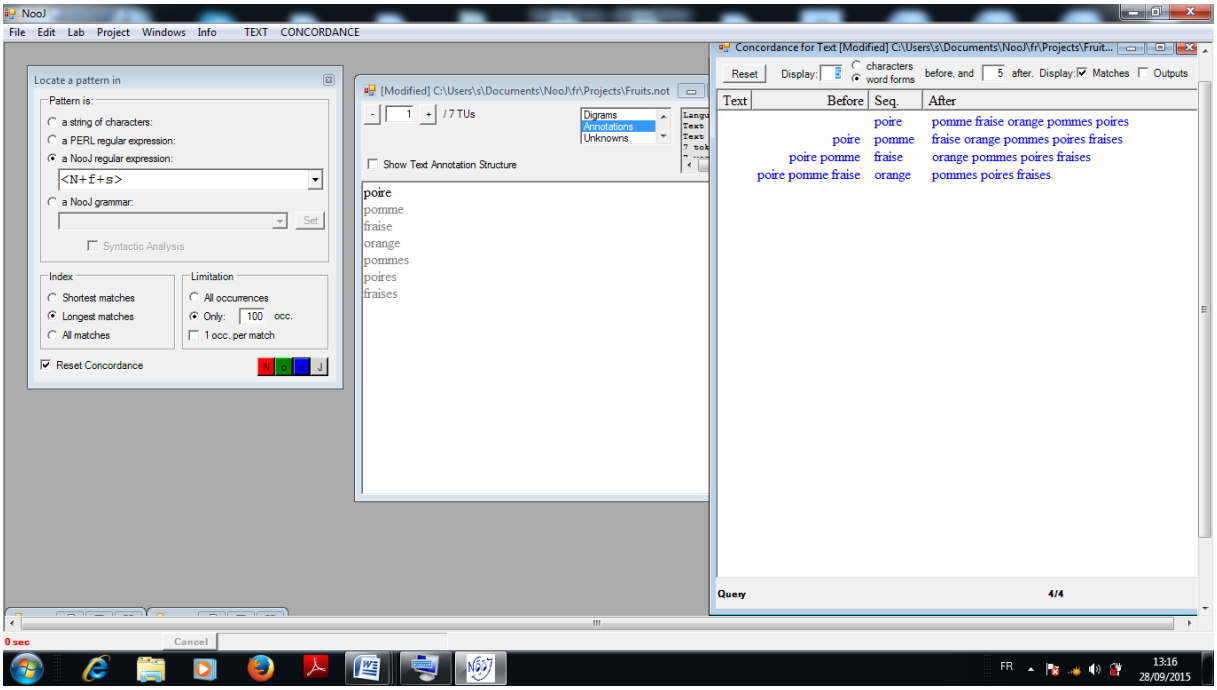


**Figure IV.7.3 : Recherche d'un patron dans le texte "fruits.not".**

**IV.7.4 Morphologie :** Pour mettre en place les règles de flexion des mots, on utilise >Lab >morphology :

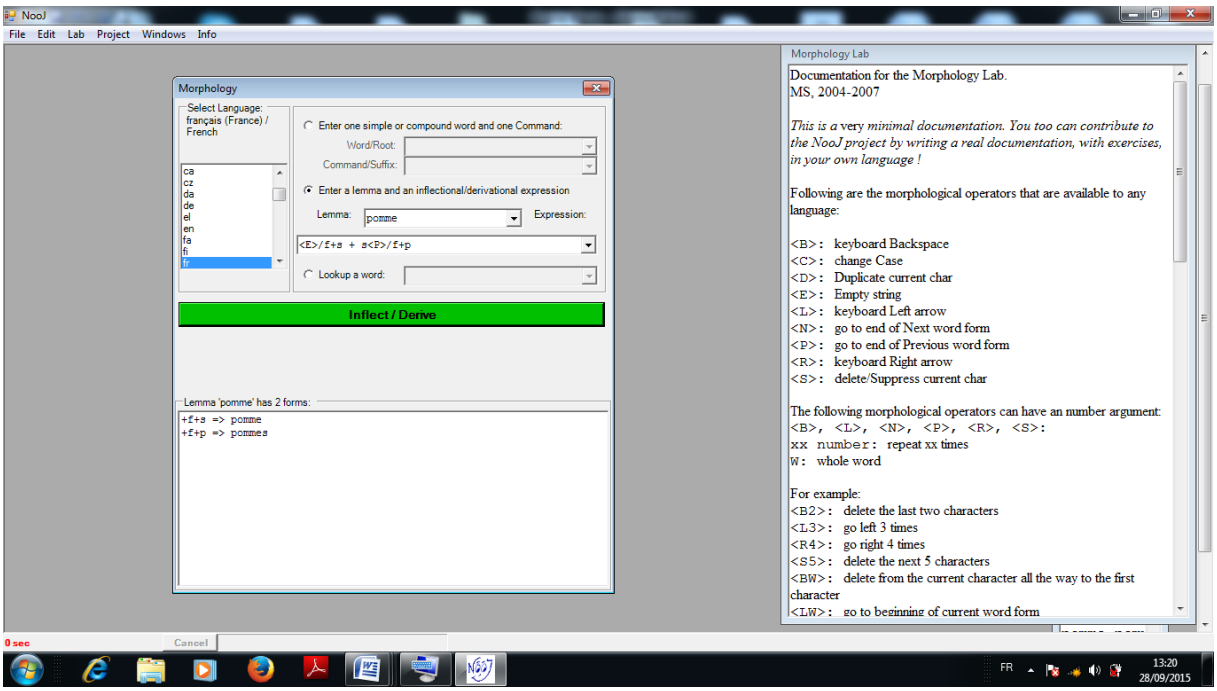


**Figure IV.7.4 : Mise au point des règles de flexion à l'aide de "Lab".**

En entrant un lemme et une expression flexionnelle ou dérivationnelle, on obtient après avoir cliqué sur Inflect/Derive, la forme des lemmes produits grâce à l'expression.

Dans la **Figure IV.7.4**, nous avons le lemme "pomme" et l'expression flexionnelle

<E>/f+s + s<P>/f+p

<E>/f+s indique qu'on ne rajoute rien "<E>" au féminin "f" singulier "s", nous obtenons :

+f+s => pomme au féminin singulier

"s<P>/f+p ", <P> indique aller à la fin du mot précédent, rajouter s, nous obtenons

+f+p => pommes au féminin pluriel.

## IV.8 Conclusion

Après avoir parcouru ce chapitre, on peut déduire que le système Nooj représente un outil puissant. Avec une plate-forme conviviale, il permet le développement et la mise au point de diverses applications, linguistiques, d'aide à l'enseignement des langues, d'aide à la traduction automatique et enfin d'extraction de terminologie.

# *Chapitre V*

# *Ressources linguistiques et Exemples*

# Chapitre V:Ressources linguistiques et Exemples [13]

## V.1Introduction

Après avoir décrit dans le chapitre précédent notre environnement de travail NooJ, nous procéderons dans celui-ci à la présentation de notre collaboration qui a pour but de réaliser une traduction automatique des termes de l'informatique.

## V.2Dictionnaire bilingue de traduction Français- Anglais

### V.2.1 extrait du dictionnaire

La figure ci-dessous représente un extrait du dictionnaire français-anglais, le mot en français est suivi de son équivalent en anglais. On a utilisé pour sa construction le dictionnaire Hild [13]

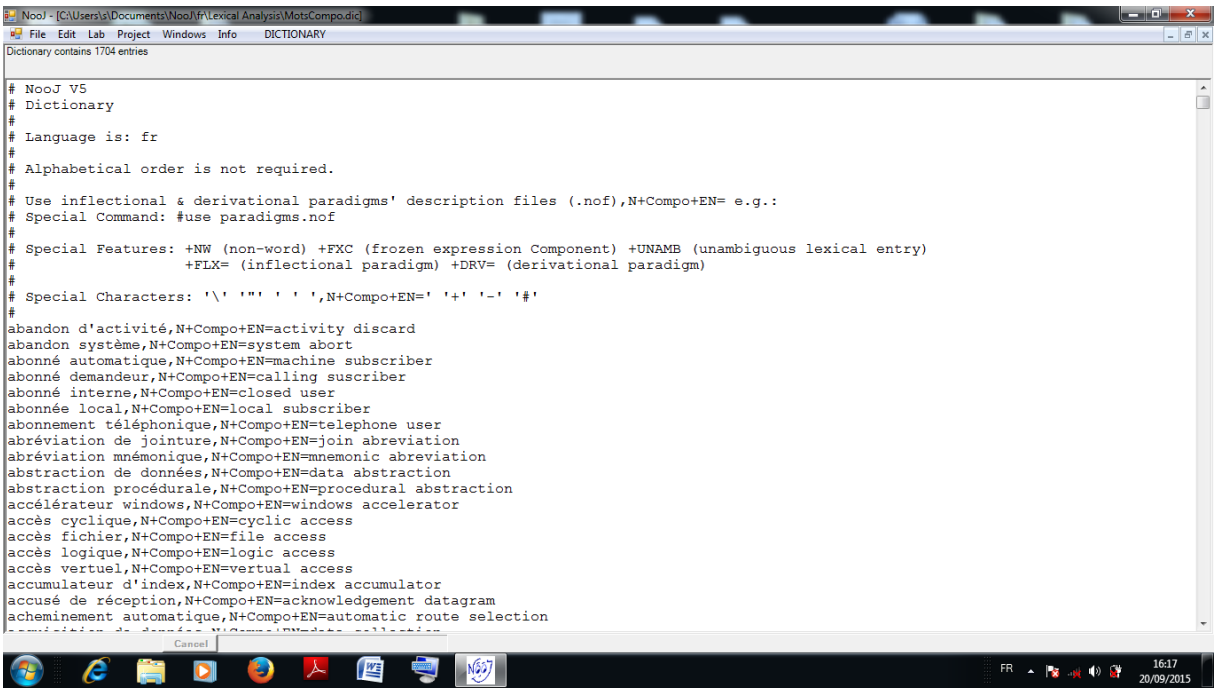


**Figure V.2.1 extrait du dictionnaire français-anglais**

### V.2.2 format d'une entrée

On prend comme entrée le mot « abandon d'activité » **(Figure V.2.2):**

abandon d'activité,NDN+Compo+EN=activity discard

Le terme ici appartient à la classe des noms composés Nom1 de Nom2 (NDN).'Compo' indique que le terme est composé, en effet, il est constitué de « abandon » et « activité ».

[13] Jacques Hildebert,dictionnaire des technologies de l'informatique,vol.2, Français-Anglais, la maison du dictionnaire(Paris), Hippocrene Books Inc, (New York),1998.

‘EN’ indique qu’on aura en sortie l’équivalent du terme en anglais après l’avoir reconnu.

## V.3 Dictionnaire bilingue Anglais-Français

### V.3.1Extrait du dictionnaire

La figure ci-dessous représente un extrait du dictionnaire anglais-français, le mot est en langue anglaise suivi de son équivalent en langue française. Ici on construit le dictionnaire à partir d’un aide-mémoire [14]

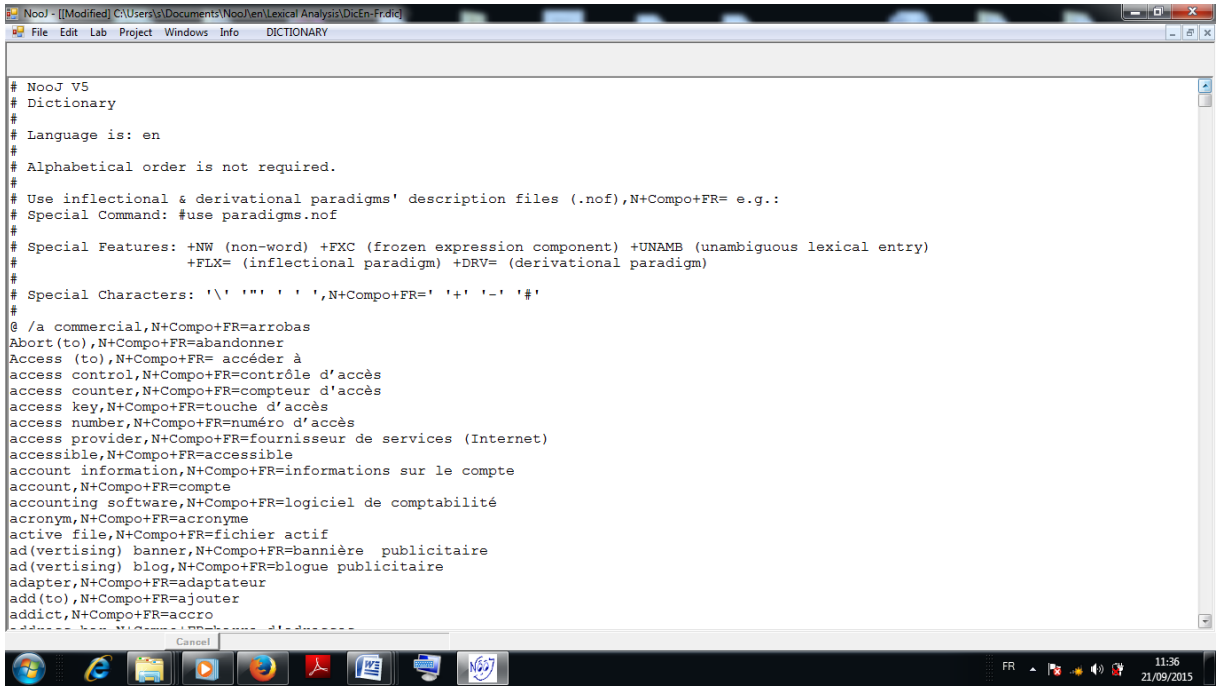


**Figure V.3.1 extrait du dictionnaire anglais-français**

### V.3.2 Format d’une entrée

On prend cette fois-ci comme entrée le terme « access control » :

access control,N+Compo+FR=contrôle d’accès

Comme en français, le terme est composé de deux mots simples « access » et « control » comme l’indique le ‘Compo’. Le ‘N’ indique que c’est un nom. ‘FR’ indique qu’on aura en sortie l’équivalent du terme en français après l’avoir reconnu.

[14] Aide mémoire Anglais-Français des termes de l’informatique, de l’internet et de l’editique (Enrichi de co-occurrences usuelles. MYOSOTIS Presse.

## V.4Grammaires syntaxiques de traduction

### V.4.1 Traduction sans flexions

Dans ce cas, la traduction des mots se fait sans les règles de flexions, on utilisera la grammaire suivante pour la traduction du français vers l'anglais.

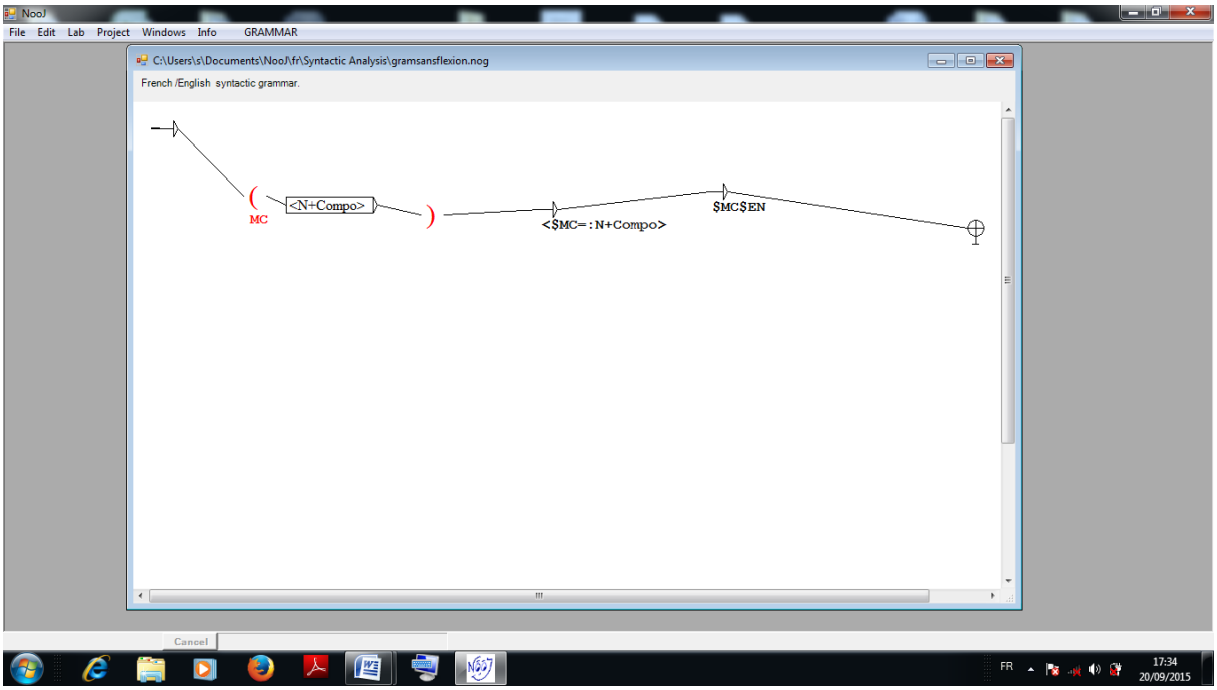


**Figure V.4.1(a) Grammaire syntaxique français-anglais sans flexions**

Et celle-ci pour la traduction de l'anglais vers le français **(Figure V.4.1(b))**

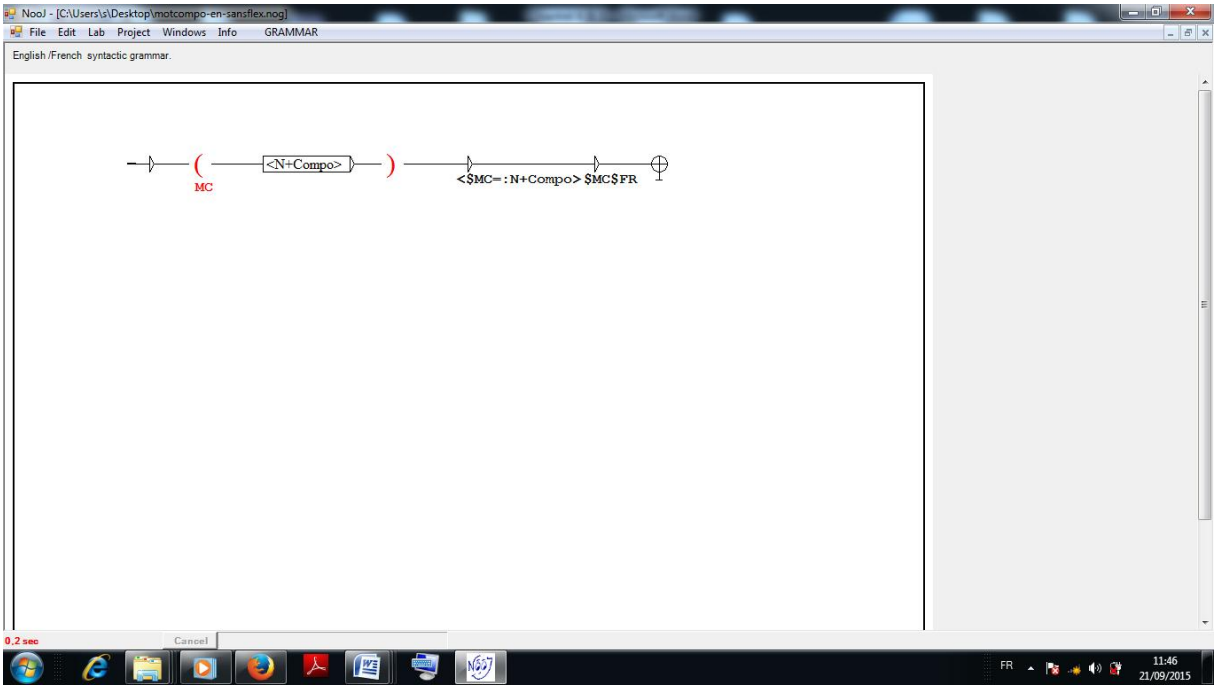


**Figure V.4.1(b) Grammaire syntaxique anglais-français sans flexions**

### V.4.2 Traduction avec flexions

Dans ce cas la traduction des mots se fait avec une grammaire flexionnelle, comprenant des règles de flexions des mots à fléchir.

### V.4.2.1 dictionnaire des mots composés

Pour reconnaitre et traduire toutes les formes du terme composé, il faut rajouter une autre ressource linguistique qui est la grammaire des flexions.

Dans l'entrée du dictionnaire, il faut rajouter la propriété +FLX=xx, où xx est le paradigme (règle) de flexion.

Il faut également mentionner la commande #use 'le nom de la grammaire flexionnelle'.nof pour indiquer la grammaire de flexions utilisée afin de reconnaitre les formes fléchies.

### V.4.2.2 Exemple d'une entrée

La **figure V.4.2.2(a)** montre le format de quelques entrées en français, et la **figure V.4.2.2(b)** le format de quelques entrées en anglais

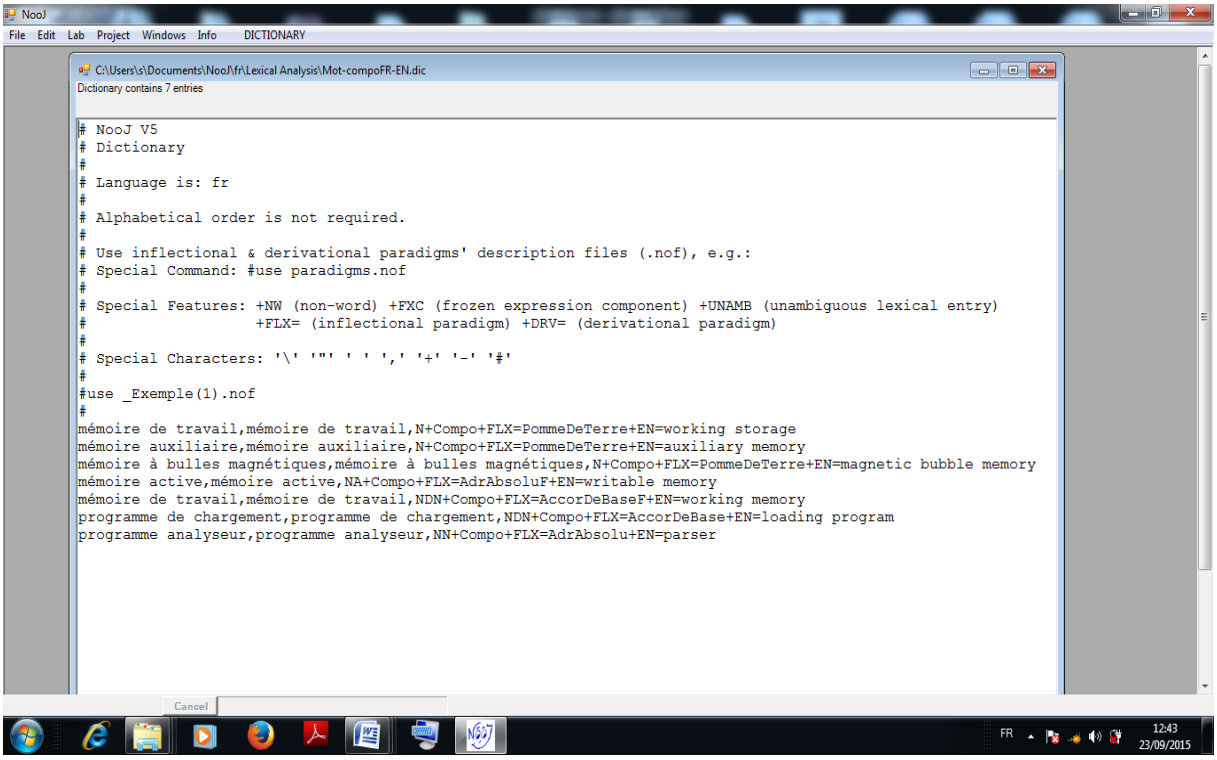


**Figure V.4.2.2(a) format d'entrée (FR)**

Prenons comme exemple :

Mémoire active,mémoire active,NA+Compo+FLX=AdrAbsoluF+EN=writable memory '+FLX=AdrAbsoluF' permettra de reconnaitre 'mémoire active' au féminin singulier et mémoires actives au pluriel comme on le verra par la suite.

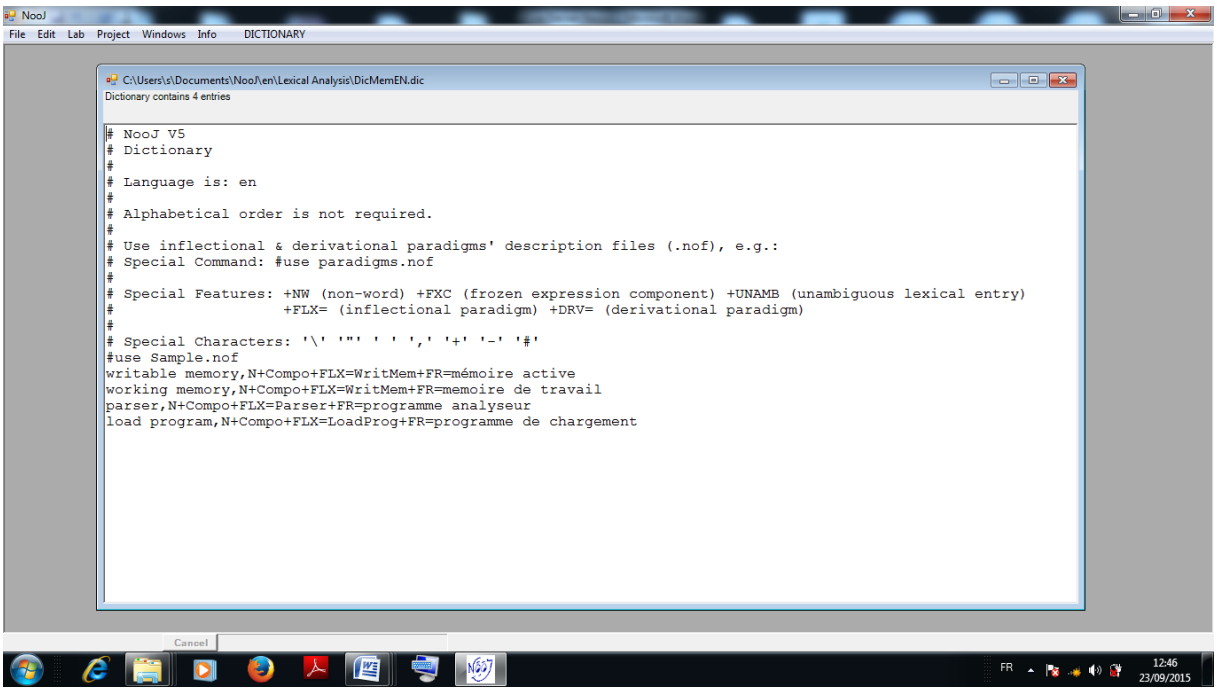


**Figure V.4.2.2(b) format d'entrée (EN)**

Prenons comme exemple :
Working memory,N+Compo+FLX=WriteMem+FR=memoire de travail
Ici '+FLX=WriteMem' permettra de reconnaitre 'write memory' au singulier et 'write memories' au pluriel.

### V.4.2.3 Grammaire flexionnelles

#### V.4.2.3.1 Termes en français

Les règles de flexions sont regroupées dans un fichier '_Exemple(1).nof' La **figure V.4.2.3.1** ci-dessous, présente un extrait du fichier.

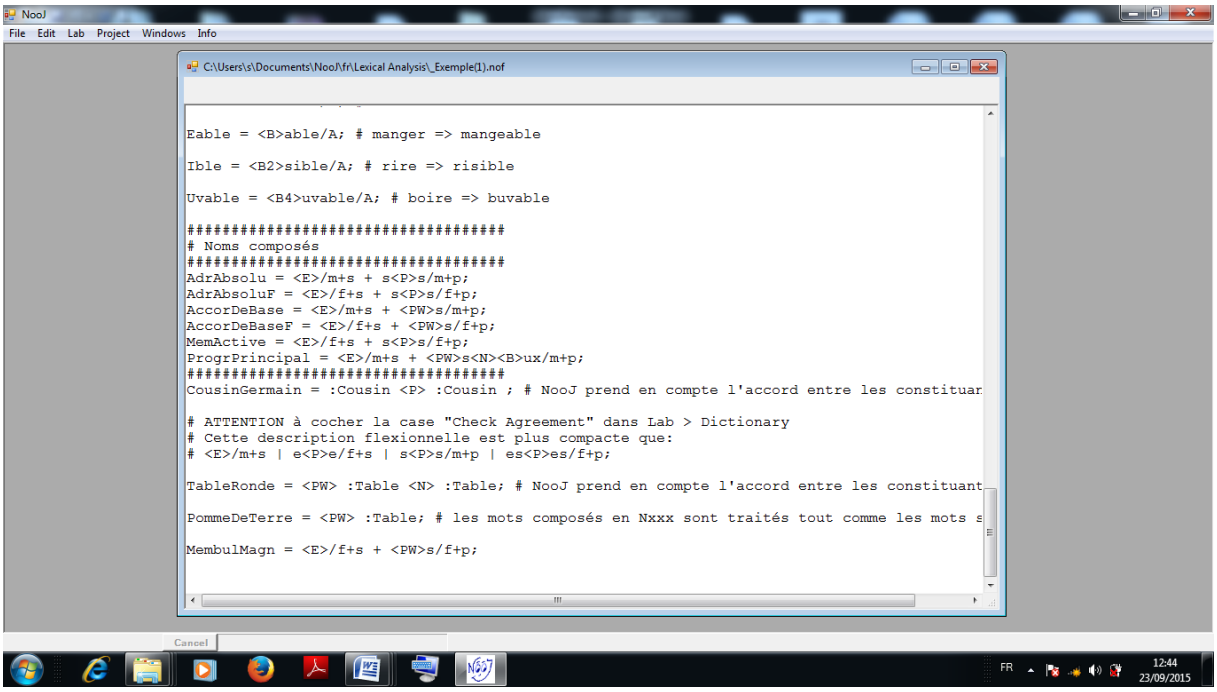


**Figure V.4.2.3.1(a) Règles de flexions(FR)**

La figure ci-après montre les mots fléchis correspondants :

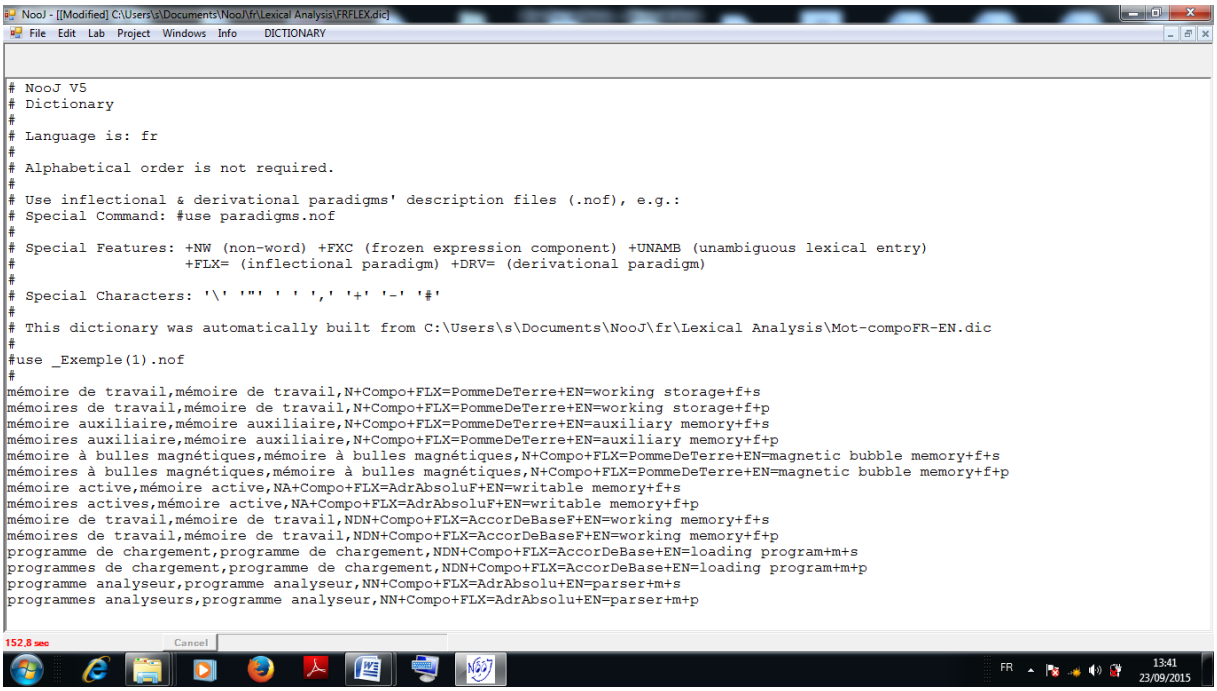


**Figure V.4.2.3.1(b) Mots fléchis(FR)**

### V.4.2.3.2 Termes en anglais

De même les règles de flexions sont regroupées dans le fichier nommé cette fois ‘Sample.nof’

La **figure V.4.2.3.2(a)** présente un extrait du fichier.

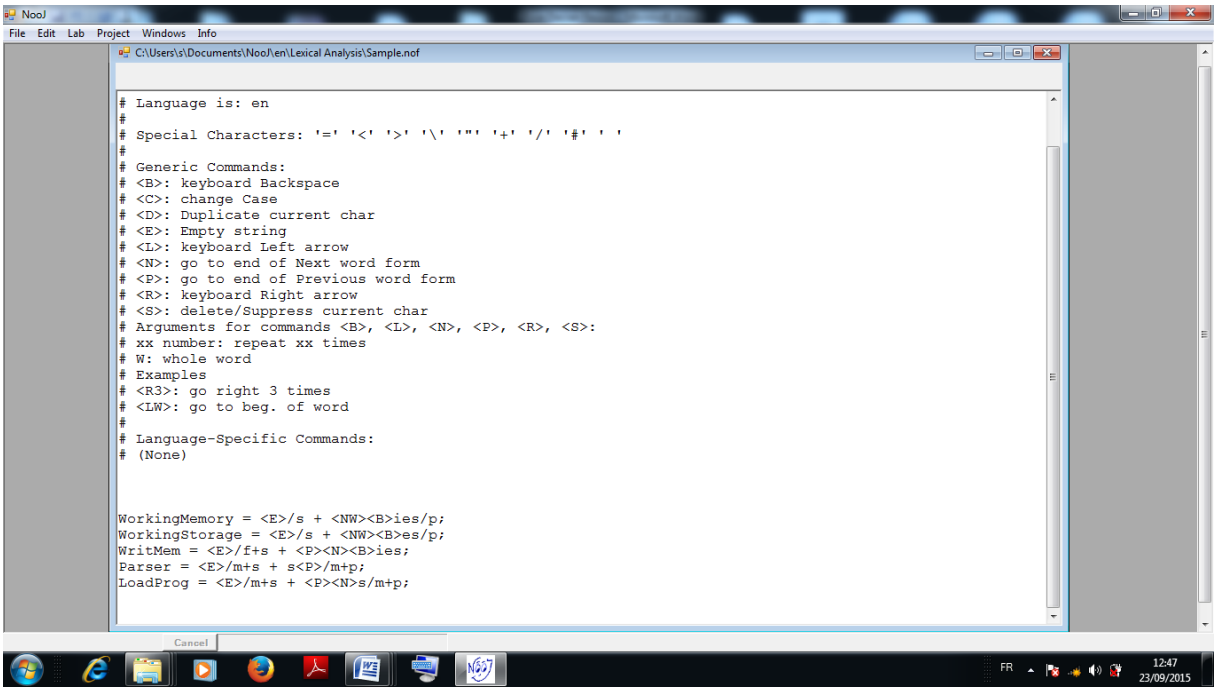


**Figure V.4.2.3.2(a) Règles de flexions (EN)**

La figure ci-après présente les mots fléchis correspondants.

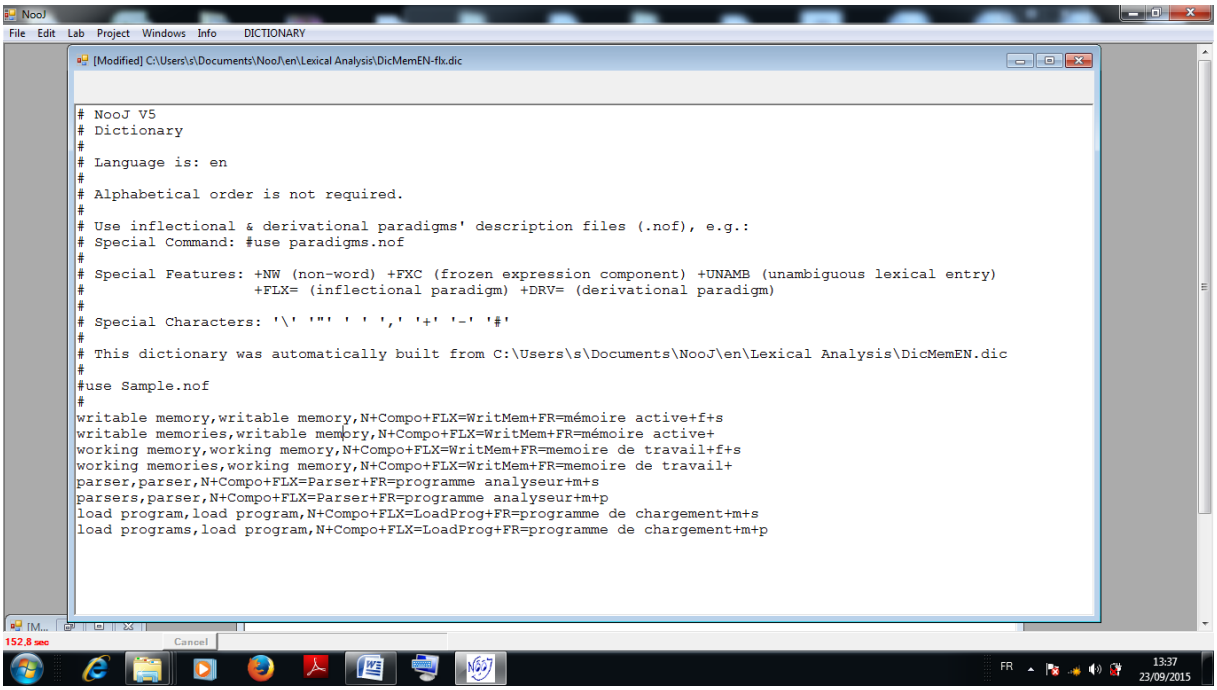


**Figure V.4.2.3.2(b) Mots fléchis(EN)**

### V.4.2.4 Grammaire syntaxique avec flexions

Dans ce cas, la traduction des mots se fait en prenant en considération les règles de flexions, on utilisera la grammaire suivante pour la traduction du français vers l'anglais.

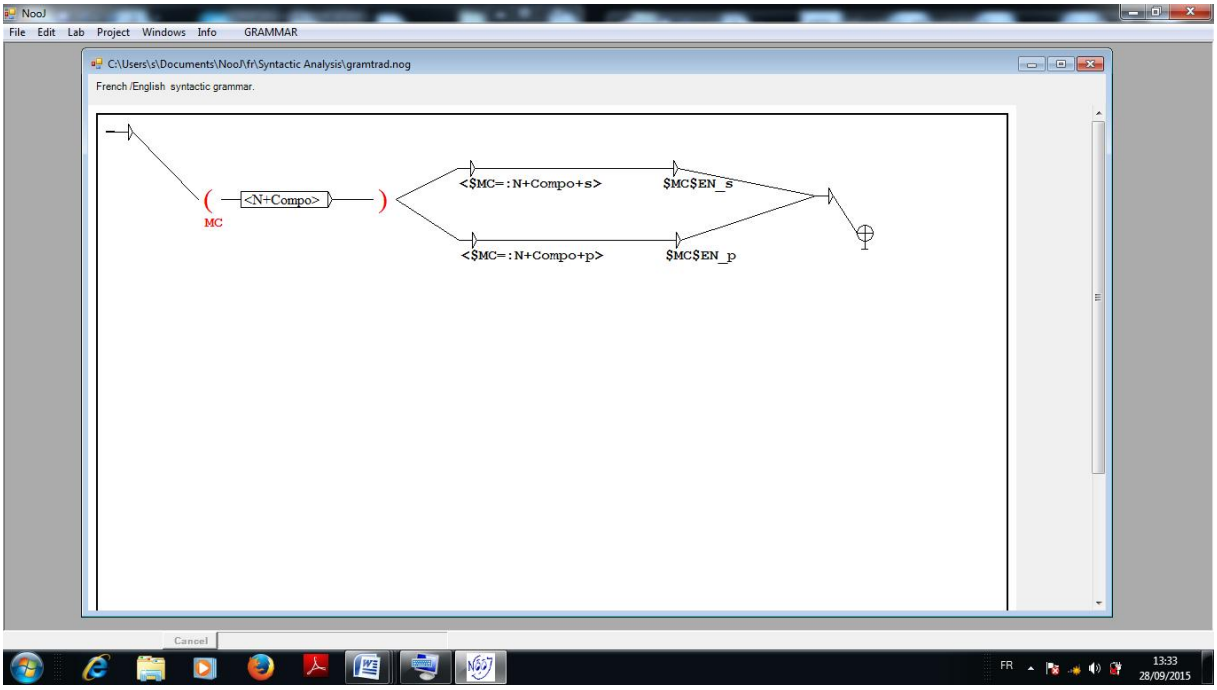


**Figure V.4.2.4 Grammaire syntaxique avec flexions**

## V.5 Texte

Dans cette partie on présente les textes choisis pour tester la traduction, ils contiendront essentiellement des termes d'informatiques.

### V.5.1 texte en français

Le fichier 'TexteFRANCAIS.not'contient des termes composés allant de la lettre A à Z. Dans la figure V.5.1 nous trouverons un extrait du fichier.

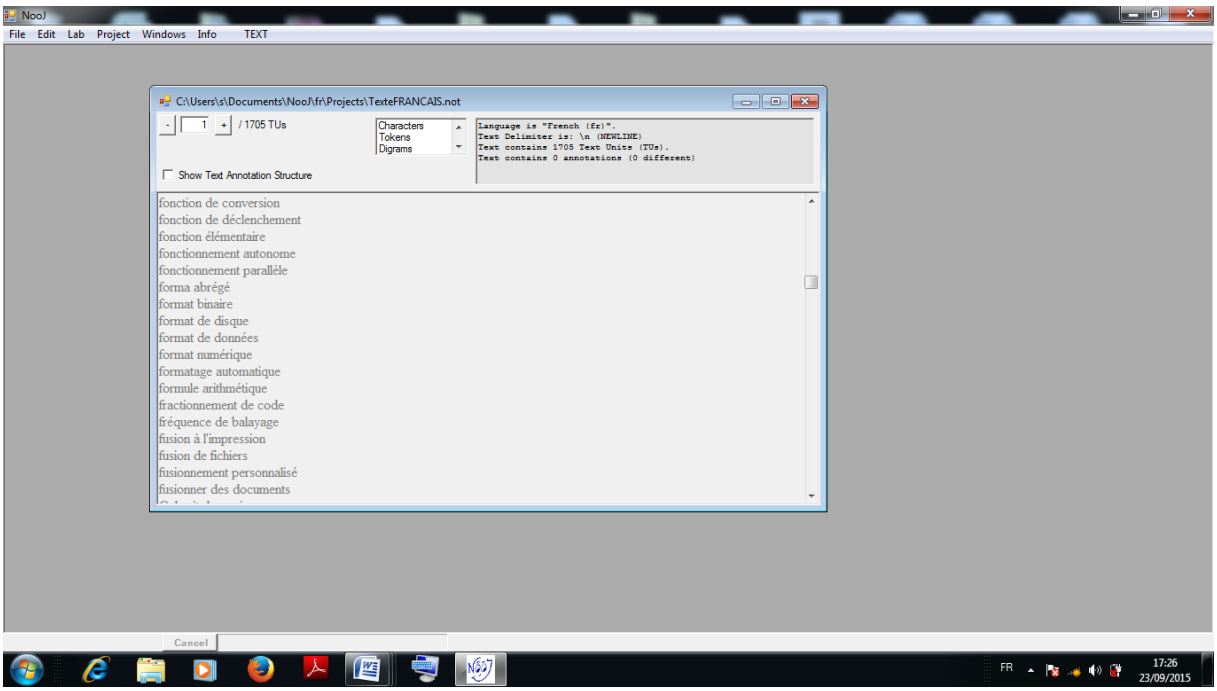


**Figure V.5.1 Extrait du fichier 'TexteFRANCAIS.not'**

### V.5.2 Texte en anglais

La figure suivante (**Figure V.5.2**) présente un extrait du fichier 'TextANGLAIS.not'qui comporte des termes simples et composés en anglais allant de la lettre A jusqu'à la lettre Z.

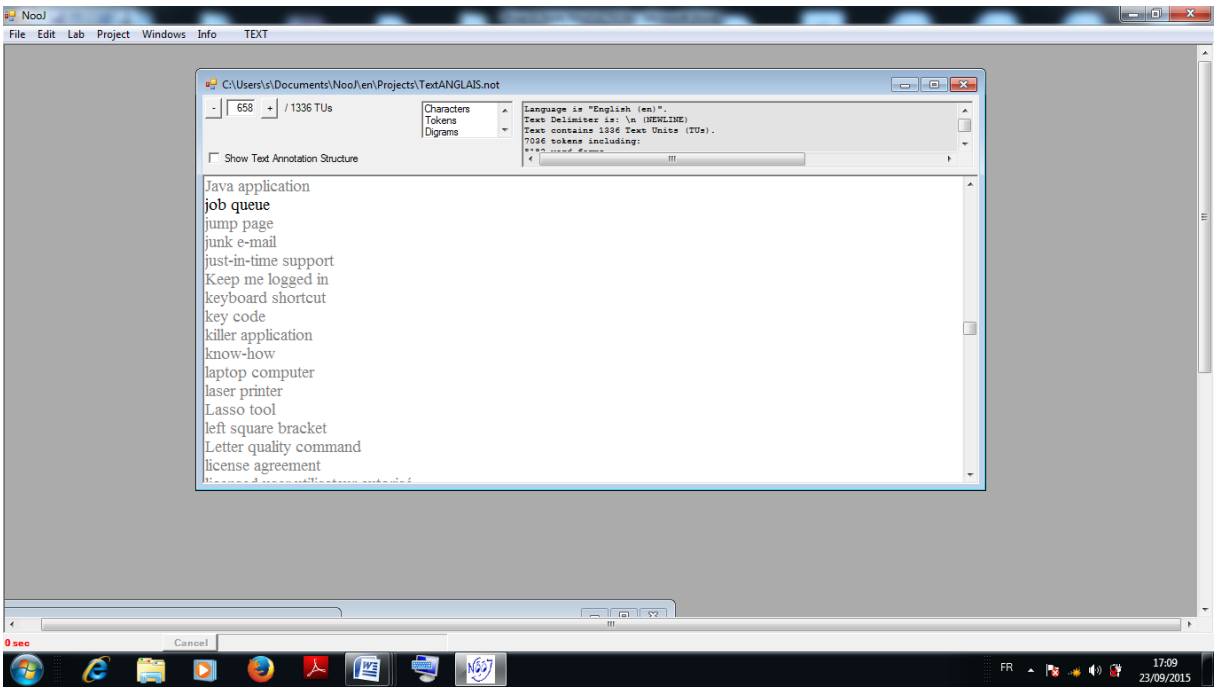


**Figure V.5.2 Extrait du fichier 'TextANGLAIS.not'**

## V.6 Résultats de traduction

### V.6.1 FR-EN

Dans la figure suivante, on trouvera un extrait du résultat de la traduction du texte contenue dans le fichier 'TexteFRANCAIS.not' (**Voir Figure V.5.1**)

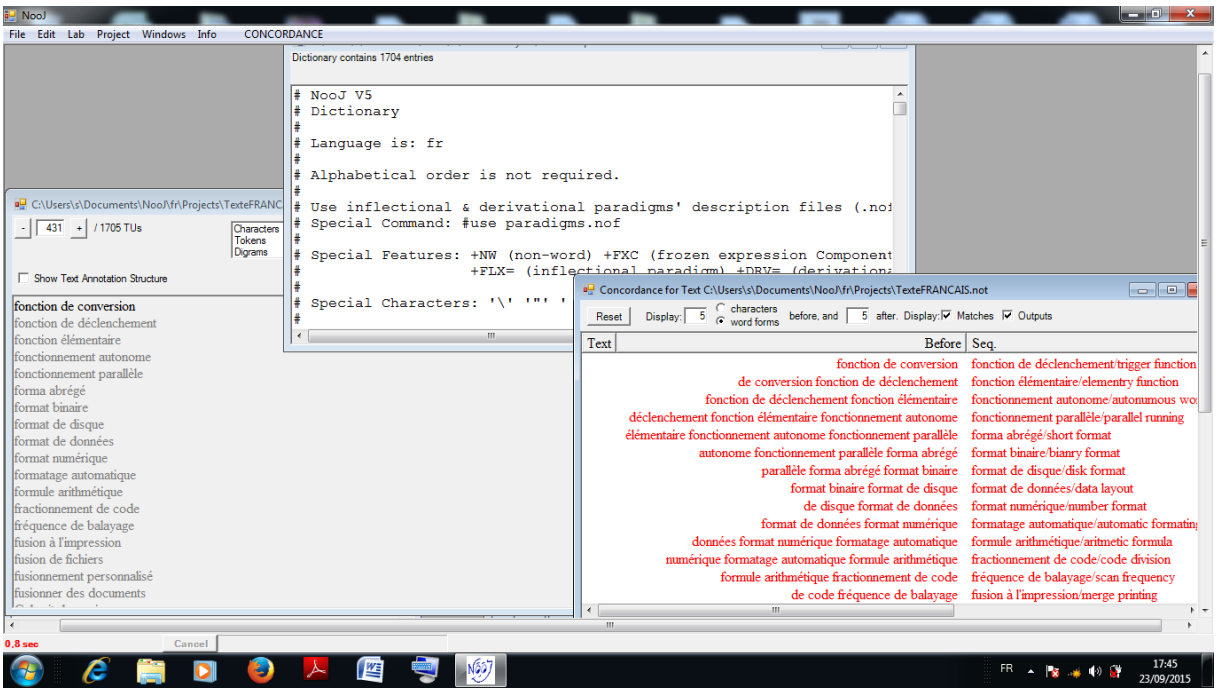


**Figure V.6.1 Exemple de traduction (FR-EN)**

### V.6.2 EN-FR

Dans la figure suivante, on trouvera un extrait du résultat de la traduction du texte contenue dans le fichier 'TextANGLAIS.not' (**Voir Figure V.5.2**)

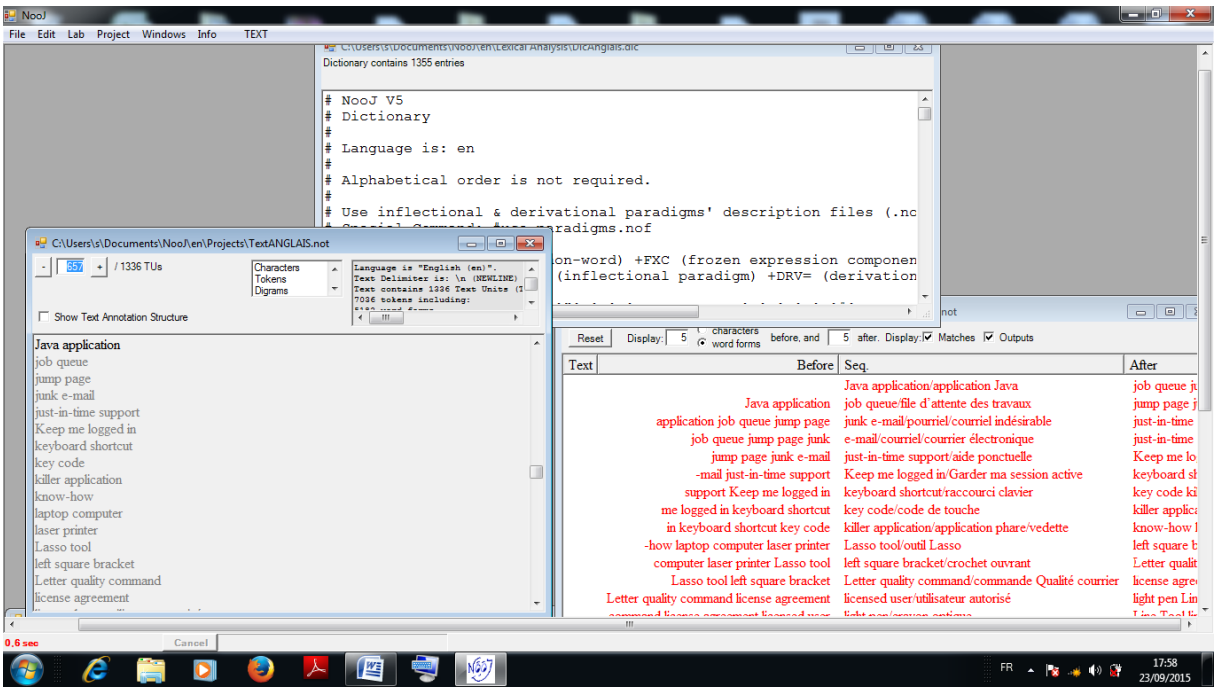


**Figure V.6.2 Exemple de traduction (EN-FR)**

## V.7 Conclusion

Dans ce dernier chapitre, nous avons pu présenter les différentes ressources linguistiques qu'on a pu construire à l'aide de Nooj, le dictionnaire construit ainsi que les grammaires syntaxiques de traduction, mais aussi quelques règles de flexion.

# *Conclusion et perspectives*

# Conclusion et perspectives

Comme pour tous les autres nouveaux domaines de la découverte humaine et du progrès technologique, la traduction automatique a connu diverses périodes. Les chercheurs sont passés de l'optimisme délirant et de l'enthousiasme, au désespoir noir et à la désillusion. Il a fallu l'acharnement de quelques-uns, qui avaient vraiment foi, et l'aide d'institutions financièrement puissantes pour persévérer, modifier et développer les systèmes avec obstination.

Le but de notre travail a été de développer des ressources linguistiques en vue de la traduction automatique de terminologie informatique.

Le cadre de recherche est le système NooJ. Nous avons mis en place deux dictionnaires français-anglais, séparés, celui des termes simples, et celui des mots composés. Egalement deux dictionnaires anglais-français contenant respectivement les mots simples, les mots composés.

Nous avons aussi construit la grammaire locale de traduction (sans flexion) et également la grammaire flexionnelle contenant quelques règles de flexion des termes composés du français et celle pour les termes de l'anglais.

A travers ce projet, nous avons pu nous familiariser avec le domaine du traitement automatique des langues, qui est d'ailleurs très vaste. L'environnement de développement linguistique Nooj, nous avons appris également divers concepts portant sur la traduction automatique, et enrichis notre vocabulaire informatique.

Les principales perspectives de ce travail consistent en :

- La traduction de textes techniques portant sur l'informatique.
- L'ajout d'autres termes afin de compléter les dictionnaires déjà établis.
- La construction de nouvelles flexions pour les autres termes français et anglais composés.
- La traduction avec flexions.

# *Bibliographie*

## Bibliographie

[1] BOUILLON,Pierette. « *Traitement automatique des langues naturelles* ». *Champs Linguistiques.Champs linguistiques: Recueils*, ISSN 1377-7971 *Universités francophones*, ISSN 0993-3948. Edition Duculot AUPELF UREF. Bruxelles. De Boeck Supérieur, 1998, 2801111813, 9782801111819. 245pages.

[2] JACQUEMIN.Christian,ZWEIGENBAUM.Pierre. « *Traitement automatique des langues pour l'accès au contenu des documents* ». [Document électronique]. Paris.Anass.2007.
https://perso.limsi.fr/pz/FTPapiers/Jacquemin:I32000.pdf

[4] POUDAT Céline. « Outils de traitement de corpus ». [Document électronique]. Université d'Orléans.juin-septembre 2003.
http://www.revue-texto.net/Corpus/Manufacture/pub/Poudat_Outils.html

[5]YVON,François. « *Une petite introduction au Traitement Automatique des Langues Naturelles* ».[document électronique].France,2007,
http://formation.enst.fr/INF343/polyTLN.pdf

[6] SARADOUNI.Y,TITEM.A,MESSAOUDENE.N.Mémoire de fin d'études. Dirigé par AOUGHLIS Farida. « *Transformation des requêtes, du langage naturel à un script d'interrogation en langage métriques* ». Université UMMTO.Tizi-Ouzou.2007/2008.

[7] KOUASSI, Roland. «*La problématique de la traduction automatique* ». [Document électronique]. Université de Cocody.
www.ltml.ci/files/articles4/article_traduction_automatique.pdf

[8] *GRASS, Thierry. « A quoi sert encore la traduction automatique ? » N°2/2010. Outils de traduction – outils du traducteur. Groupe d'étude sur le plurilinguisme Européen.*
http://www.cahiersdugepe.fr/index1367.php

[9] PAVEL,Sylvia. NOLET,Diane. «Précis de terminologie». [Document électronique]. Bureau de la traduction, Travaux publics et services gouvernementaux Canada.
www.termsciences.fr/sites/.../IMG/pdf/precis_de_terminologie_Pavel.pdf

*[10] AOUGHLIS YAMOUNI Farida. Construction d'un dictionnaire électronique de terminologie informatique et analyse automatique de textes par grammaires locales. Thèse de doctorat d'état en informatique soutenue le 12/12/2010, Université Mouloud Mammeri Tizi-Oouzou (UMMTO).*

*[11] Cécile Yapaudjian-Labat « Quelques éléments théoriques autour du lexique ». [Document électronique].*
http://www.ac-orleans-tours.fr/fileadmin/user_upload/ia28/doc_peda/MDL/outils/vocabulaire/lexique-elements-theoriques.pdf

[12] Costăchescu Adriana. Comment créer une terminologie ? (Lexique de l'informatique). Faculté des lettres, Université de Craiova (Roumanie). Synergies *Royaume-Uni et Irlande* n° 5 - 2012 pp. 141-155.

[13] YAMOUNI Farida. « A French-Tamazight MT system for Computer Science Compound Words» NOOJ'15, 11-13 june 2015, Minsk, Belarus

## Réferences

Fondation Rockefeller : fondation caritative privée, dotée du statut fiscal 501c3, fondée par John Davison Rockefeller et Frederick T. Gates, destinée à promouvoir le progrès scientifique dans tous les pays du monde

Allen1987: Jerry and Allen, Janice. 1987. Halia Grammar. (Data Papers on Papua New Guinea Languages, 32.) Ukarumpa, Papua New Guinea: Summer Institute of Linguistics.

Sabah 1988 : Gérard Sabah, L'intelligence artificielle et le langage. Représentation des connaissances, Hermès, Paris, 1988

Weizebaum 1966 et 1967 :Joseph Weinzenbaum, ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine, Communications of the ACM, janvier 1966

Winograd: Winograd,Terry, Understanding Natural Language, (191 pp.) New York: Academic Press, 1972. Also published in Cognitive Psychology, 3:1 (1972), pp. 1-191.

Woods, Willian A., 1968 Procedural semantics for a question –asnwering machine. AFIPS Conference Proceedings 33 (FJCC), 457-71

MUC,MUC6 Message Understanding Conference (MUC) 6 a été produit par Linguistic Data Consortium (LDC) le numéro de catalogue LDC2003T13 and ISBN 1-58563-239-2. En 1990, les évaluations MUC a financé le développement de mesures et des algorithmes

statistiques pour appuyer des évaluations gouvernementales des nouvelles technologies d'extraction d'information

HUTCHINS John, 2004, « Machine translation and computer-based translation tools. A new spectrum of translation studies », dans BRAVO José Maria (ed), Publicationes de la universidad de Valladolid, p. 13-48.

LOFFLER-LAURIAN Anne-Marie, 1983, « Pour une typologie des erreurs dans la traduction automatique », dans *Multilingua*, vol. 2, n° 2, p. 65-78.

GROSS Gaston, 1995, « Une sémantique nouvelle pour la traduction automatique : les classes d'objets », dans *La Tribune des Industries de la Langue et de l'Information électronique*, n°17-18-19, pp. 16-19.

LAKOFF George, 1972, « Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts », dans PERANTEAN P. M., LEVI J. N., and PHARES G. C. (ed.), *Papers from the 8th Regional Meeting*, Chicago Linguistics Society, p. 183-228.

MEL'ČUK Igor, 1998, « Collocations and Lexical Functions », dans COWIE Anthony P. (ed.), *Phraseology: Theory, Analysis and Applications*, (Oxford Studies in Lexicographie and Lexicology), Oxford, Oxford University Press, p. 79-100.

Benveniste (1966,1967) problèmes de linguistique générale.Paris.PLG vol1 (1996), PLG vol2 (1967)

Gross M.1986. Grammaire trasformartionnelle du Français. Rapport technique n°3, Paris, Programme de recherches coordonnées « informatique linguistique », CNRS.

Piot M.1988. Conjonctions de subordination et figement. Langages, vol.90, 1988. Larousse, Paris.

Danlos L.1980, Représentation d'informations linguistiques : les constructions N être PrepX.Thèse de PhD, université de Paris 7

Gross M.1989. Electronic Dictionaries and Automata in Computational Linguistics, Lecture Notes in Compter Science, Berlin : Springer Verlag, 110p. Gross Maurice, Dominique Perrin eds.1989

Daille B.1994.Approche mixte pour l'extraction automatique de terminologie: statistique lexical et filters linguistiques, Thèse de doctorat en informatique fondamentale, université de Paris7.

Monceaux A.1993. la formation des noms composés de structure Nom Adj, élaboration d'un lexique électronique, Thèse de Doctorat, université de MLV, Noisy le Grand.

Mathieu-Colas M.1988. Typologie des noms composés. Rapport technique du « Programme de Recherches Coordonnées Informatique Linguistique »,C.N.R.S

Gross G. 1990. Définition des noms composés dans un lexique-grammaire, dans Langue Française, septembre 1990, N°87

Gross G. 1991. Analyse groupes N de N. Dans Actes du colloque des industries de la langue.Nantes

Gross M.1986. Grammaire transformationnelle du Français, 2- Syntaxe du nom, éditions Cantilène.

Domingues C.2001.Etude d'outils informatiques et linguistiques pour l'aide à la recherche automatique d'information dans un corpus documentaire. Thèse de docteur de l'université de Marne La Vallée.

Humbley, 1974 : « vers une typologie de l'emprunt linguistique ».cahiers de lexicologie, 25,pp.46-70

Koch, Peter.2002. « lexical typology from a cognitive and linguistic point of view". In Alan D. cruse et al.(eds.), lexikologie: ein internationals Handbuch zur Natur und Struktur Von Wôrtern und wortschatzenn/ lexicology: an international handbook on the nature and structure of words and vocabularies. Berlin/ New York: walter de Gruyter: 1142-78

Stanforth, Anthony W.2002. "effects of language contact on the vocabulary: an overview". In Ala D.Cruse et al.(eds.), lexikologie: ein internationals handbuch zur natur und struktur von wortern und wortschatzen/ lexicology: an international handbook on the nature and structure of words and vocabularies. Berlin/new York: walter de gruyter: 805-13

Thibault, André.2004. Evolutions sémantiques et emprunts: le cas des gallicismes de l'espagnol

Trotter, David.2009. « l'apport de l'anglo-normand à la lexicographie de l'anglais, ou : les 'gallicismes ' en anglais ». In andré Thibault (éd), Gallicismes et théorie de l'emprunt linguistique. Paris : l'Harmattan : 147-68.

Iliescu, Maria, Costachescu, Adrina, Dinca,Daniela, Popescu, Mihaela, Gabriela.2010. « typologie des emprunts lexicaux français en roumain ». Revue de linguistique Romane, 75,pp 589-604.

Montaigne 1595 : Montaigne, Essais, édition de 1595, avec ajouts manuscrits (couches B et C de l'édition de Bordeaux)

Horn, Laurence Ward, Gregory (eds.).2004. The Handbook of Pragmatics. Oxford: Blackwell Publishing.

Silberztein M. et Tutin A. 2005 NooJ: Un outil Tal de corpus pour l'enseignement des langues. Application pour l'étude de la morphologie lecicale en FLE, Article n°alsic_v08_20-rec11,p.123-124