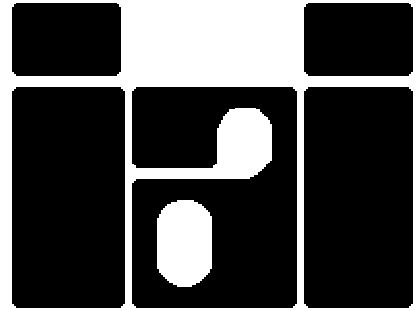


**Université de Metz**  
**UFR Lettres et Langues**



**Institut für**  
**Angewandte**  
**Informationsforschung**

# Traitement Automatique des Langues et Terminologie



## **Remerciements**

Je tiens à remercier toutes les personnes grâce auxquelles ce stage a pu avoir lieu, et ce de plus dans d'excellentes conditions.

Ceci s'adresse en particulier au Professeur Johann Haller et au Docteur Axel Theofilidis, mais également à toute la compétente et sympathique équipe de l'IAI.

# Sommaire

<b>INTRODUCTION</b>	<b>2</b>
<b>1 TÂCHES EFFECTUÉES</b>	<b>3</b>
1.1 A LA DÉCOUVERTE DE VI	3
1.1.1 <i>Correction de dictionnaire</i>	3
1.1.2 <i>Les inconnus de l'analyse morphologique</i>	3
1.2 TRADUCTION	4
1.2.1 <i>L'IAI en français sur le Net</i>	4
1.2.2 <i>Localisation : adaptation en français d'un logiciel</i>	4
1.3 TERMINOLOGIE	7
1.3.1 <i>Création d'une liste de termes multilingue</i>	7
CRÉATION D'UNE LISTE BILINGUE DE TERMES À PARTIR DE TEXTES ALIGNÉS	8
<b>2 DIAGNOSTIC D'UNE BASE DE DONNÉES TERMINOLOGIQUES EXISTANTE</b>	<b>11</b>
2.1 SITUATION DE DÉPART	11
2.1.1 <i>Données de Boehringer Ingelheim</i>	11
2.1.2 <i>Ressources de l'IAI</i>	12
2.2 ANALYSE PRÉALABLE DE LA BASE DE DONNÉES	12
ANALYSE DE LA BASE DE DONNÉES TERMINOLOGIQUE DE BOEHRINGER INGELHEIM <i>DIAGRAMME DE PRINCIPE</i>	13
2.3 ANALYSE DES TERMES	14
2.3.1 <i>Orthographe</i>	14
2.3.2 <i>Cohérence interne</i>	15
2.4 RÉSULTATS	16
<b>3 CRÉATION D'UNE BASE DE DONNÉES TERMINOLOGIQUES</b>	<b>18</b>
3.1 STRUCTURE D'UNE FICHE TERMINOLOGIQUE	19
3.1.1 <i>Définitions</i>	19
3.1.2 <i>Catégories de données</i>	19
3.2 DE LA BASE DE DONNÉES AU SYSTÈME NOTIONNEL	21
3.3 CONSIDÉRATIONS D'ORDRE PRATIQUE	22
3.3.1 <i>Nature des champs</i>	22
3.3.2 <i>Entrée des données dans les champs</i>	22
3.3.3 <i>Un exemple : la base de données terminologiques du Service de la langue française du Ministère de la Communauté française de Belgique</i>	23
<b>4 UNIX</b>	<b>25</b>
4.1 NOTIONS DE BASE	25
4.2 COMMANDES VI	25
4.2.1 <i>Passage d'un mode à l'autre</i>	25
4.2.2 <i>Déplacement du curseur</i>	26
4.2.3 <i>Sur la ligne de commande</i>	26
4.3 EXPRESSIONS RÉGULIÈRES	27
4.4 MANIPULATION DE DONNÉES EN UTILISANT DES FILTRES	27
4.4.1 <i>Fichiers de type base de données</i>	28
4.5 ENCHAÎNER DES COMMANDES	29
4.5.1 <i>Utilisation des "pipes"</i>	29
4.5.2 <i>Scripts</i>	29
<b>CONCLUSION</b>	<b>31</b>
<b>ANNEXES</b>	<b>32</b>
A. LÉGENDE DU DIAGRAMME DE PRINCIPE	32
B. RÉFÉRENCES	33

# Introduction

L'IAI, Institut de recherche appliquée dans l'information, est un institut semi-privé spécialisé dans l'ingénierie linguistique. Il a été fondé à Sarrebruck en 1985, héritier de la longue tradition de recherche dans le domaine du traitement automatique des langues naturelles à l'Université de la Sarre.

Au début de son histoire, l'IAI a notamment été responsable de la partie allemande du projet EUROTRA de l'Union Européenne.

Aujourd'hui, une quinzaine de personnes de formations diverses (informaticiens, linguistes, traducteurs) y sont employées sous la direction du Professeur J. Haller. Le fonctionnement de l'institut peut être schématisé en deux branches : projets et services.

Les projets sont menés indépendamment ou en partenariat avec des laboratoires de recherche publics ou privés. Ayant pour objet des domaines particulièrement pointus du traitement automatique des langues, le caractère à risque de ces projets est incontestable. Ils sont par conséquent financés par des subventions ("Bundesministerium für Wirtschaft") et par des partenaires industriels, et ce jusqu'au stade du prototype industriel.

Parmi les projets à visée industrielle actuellement en cours à l'IAI, citons MULTIDOC. En collaboration avec les services de rédaction technique de différents constructeurs automobiles (BMW, Volvo, Saab, Renault), l'IAI a créé une station de travail pour rédacteur technique. Une fois sa notice rédigée, le rédacteur peut la soumettre à une analyse afin de contrôler orthographe, grammaire, style, terminologie et cohérence interne du texte.

En ce qui concerne les collaborations avec d'autres laboratoires, on peut citer le projet UNL ("Universal Networking Language"). Le but du projet est de permettre la communication multilingue sur Internet. Pour cela, on cherche à utiliser un langage pivot, UNL, ainsi que deux modules permettant de passer de la langue naturelle en langage UNL et inversement : l'enconvertisseur et le déconvertisseur. L'IAI participe à la partie allemande du projet, le français est réalisé par le GETA (Professeur Boitet, Grenoble).

D'autre part, l'IAI propose également aux entreprises des services à dominante linguistique, lui permettant d'assurer une partie de son financement. Ces tâches de moindre envergure peuvent être des évaluations de systèmes de traduction automatique, des diagnostics de bases de données terminologiques ou, occasionnellement, des traductions.

Après avoir beaucoup entendu parler de l'IAI, j'ai pu faire plus amplement connaissance avec l'institut grâce aux cours que donne le Professeur Haller aux étudiants de DESS IdL à l'Université de Metz. Nous ayant signalé que l'IAI avait besoin de stagiaires, j'ai postulé et ma candidature a été acceptée pour la période du 2 mai au 27 juillet 2000. Il en a été de même pour Patrick Da Costa et nous avons effectué la plupart des travaux décrits ci-après en commun.

Ce document constitue un résumé des principaux travaux auxquels j'ai participé tout au long de mes trois mois de stage à l'IAI.

# 1 Tâches effectuées

L'équipement informatique de l'IAI comprend un poste de travail par employé : station de travail Unix ou PC.

Les PC sont réservés aux travaux ayant "un lien avec l'extérieur" : rédaction de rapports destinés aux clients, préparation de présentations PowerPoint, etc.

L'essentiel du travail linguistique et informatique est effectué sous Unix. A partir d'un PC, il est possible de se connecter aux différentes machines Unix grâce aux applications HostAccess lorsqu'un affichage textuel est suffisant ou PC-Xware pour utiliser une interface multi-fenêtrée.

Quand on est habitué à travailler sur PC, un système convivial et coloré, il n'est pas évident de se "lancer" sous Unix. L'affichage austère et la rigueur indispensable empêchent toute découverte intuitive de l'environnement. Mais une fois apprises les commandes de base pour se déplacer dans l'arborescence, (cf. 4.1), le travail à proprement parler peut commencer.

Les tâches effectuées peuvent se regrouper en trois axes :

- Unix et vi
- traduction
- terminologie.

## 1.1 A la découverte de vi

### 1.1.1 Correction de dictionnaire

Il nous a tout d'abord été demandé de corriger un dictionnaire comprenant des termes français et leur équivalent allemand. Il s'agit principalement de corriger les erreurs de frappe et de saisie. Les termes ayant trait à des domaines tels que les transports ferroviaires, la pharmacie ou le droit des médias, ils sont très spécifiques et il est le plus souvent nécessaire de vérifier leur orthographe dans des dictionnaires papier, électroniques ou encore sur des pages Internet, en utilisant un moteur de recherche tel que Yahoo!.

Ce travail de vérification, relativement répétitif, nous a permis de nous familiariser avec les commandes de l'éditeur de texte vi : déplacement du curseur, passage de mode commande en mode saisie, recherche de chaînes de caractères, effacement de caractères ou lignes etc.

### 1.1.2 Les inconnus de l'analyse morphologique

Tous les textes analysés à l'IAI sont soumis à une analyse morphologique avant tout autre traitement. Cette analyse échoue si le mot n'est pas présent dans le dictionnaire de l'analyseur ou s'il n'a pu être décomposé de façon satisfaisante. Le mot en question est alors marqué par des balises comme "unknown", inconnu. Ceci peut être le cas si le mot :

- présente une erreur de frappe ou une faute d'orthographe
- est un nom propre

- est effectivement inconnu du dictionnaire auquel fait appel l'analyste morphologique.

Les mots marqués comme inconnus sont regroupés dans un fichier qui doit être vérifié manuellement, afin de déterminer de quel type d'erreur il s'agit et, le cas échéant, d'améliorer le dictionnaire.

Cette correction est très similaire à celle du dictionnaire, à cela près qu'il s'agit ici d'une liste de mots français uniquement, et non d'entrées en français avec leur correspondant allemand.

Ce travail sur l'éditeur vi est l'étape obligée pour tous les stagiaires de l'IAI. Il représente en quelque sorte leur "baptême du feu". Il permet de se familiariser avec la philosophie Unix et d'apprendre les commandes nécessaires afin d'effectuer des tâches plus complexes. Quand les stagiaires expriment leur ennui face à cette épreuve, c'est qu'ils sont mûrs pour passer à autre chose...

## 1.2 Traduction

Etre de langue maternelle française dans une entreprise allemande peut être un avantage, surtout lorsque cette "particularité" se double d'une formation de traducteur.

### 1.2.1 L'IAI en français sur le Net

Le site Internet de l'IAI venant d'être réactualisé, nous avons traduit les pages qui n'existaient pas en français sur l'ancien site. Travailler avec des documents de format HTML nous a sensibilisé à l'utilisation des balises, ces marqueurs qui font toute la différence entre un texte et un hypertexte en codant la présentation et les liens d'une page Internet.

Dans notre cas, il fallait, sur des pages déjà mises en formes, remplacer le texte allemand par le français. Les éditeurs de pages web tels que MS FrontPage ou Netscape Composer sont plus adaptés à la création qu'à la modification, car ils possèdent un module qui modifie automatiquement les liens, et ce pas toujours à bon escient. Il est donc plus sûr d'ouvrir les documents HTML sous Unix et d'effectuer les modifications grâce à vi. Nous avons alors remplacé le texte allemand par son correspondant français, sans modifier les nombreuses balises afin de conserver la structure du texte de départ.

Nous avons également traduit les pages web du département de Traduction Automatique de l'Université de la Sarre (dirigé par le Professeur Haller) selon le même principe.

### 1.2.2 Localisation : adaptation en français d'un logiciel

L'IAI travaille activement avec l'industrie automobile. Tout d'abord avec MULTIDOC, station de travail pour le rédacteur technique, qui est utilisé lors de la phase d'assemblage du véhicule pour décrire le montage des pièces ou lors de réparations.

A l'autre extrémité de la chaîne, 8 à 9 millions de véhicules arrivent chaque année en fin de vie dans l'Union Européenne, ce qui représente 8 à 9 millions de tonnes de déchets. Le plus souvent, ces déchets sont mis en décharge et abandonnés sans autre forme de traitement, une situation qui n'a plus lieu d'être à notre époque de prise de conscience écologique.

### 1.2.2.1 Cadre légal

Dans un souci de protection de l'environnement, la Commission Européenne a publié une proposition de directive concernant les "véhicules en fin de vie" : ils devront à l'avenir être traités par les constructeurs.

Sur les 8 millions de tonnes de déchets générées chaque année par les véhicules hors d'usage (VHU), seule la fraction métallique (75%) est recyclée par des filières classiques de récupération des ferrailles. Les 25% restants sont des "résidus de broyage", un mélange de verre, céramique, plastiques, caoutchouc, mousse de sièges, textiles, peinture, huile et lubrifiants, qui est habituellement mis en décharge. Une partie de ces résidus se compose de substances dangereuses qu'il est souhaitable de traiter et de recycler.

La proposition de directive de la Commission Européenne fixe les objectifs suivants : en 2005, un minimum de 85% du poids total du véhicule devra être revalorisé ou recyclé ; en 2015, cette part devra atteindre 95% minimum.

Pour permettre de réaliser ces objectifs quantitatifs, la Commission propose des actions dans différents domaines, en collaboration avec les partenaires économiques. D'une part, les pièces rénovables dans le respect des normes de sécurité (transmissions, radiateurs, démarreurs, alternateurs, etc.) doivent être réemployées. D'autre part, un grand nombre de matériaux peuvent être recyclés et entrer dans la composition de nouvelles pièces pour l'industrie automobile. C'est notamment le cas des plastiques.

Ainsi, lors de la conception des véhicules, un marquage des pièces plastiques selon la norme ISO 11469 (ISO 11469:2000 Plastiques -- Identification générique et marquage des produits en matière plastique) facilite le tri ultérieur par famille de matériau. Mais la rentabilité du recyclage des composants en matière plastique dépend en grande partie du temps de démontage nécessaire. Les constructeurs sont invités à produire des manuels et schémas de démontage. Afin d'optimiser le désassemblage, un logiciel est d'autre part en cours de développement dans un centre allemand de recherche en informatique.

### 1.2.2.2 Pourquoi localiser ?

Un logiciel est développé en tenant compte des impératifs techniques des fonctionnalités requises. Les informaticiens conçoivent l'architecture du programme, sur laquelle vient ensuite se greffer l'interface de l'utilisateur. Celle-ci prend aujourd'hui le plus souvent la forme d'une interface graphique, avec ses menus, boutons et boîtes de dialogue. Lorsque le logiciel s'adresse à un large marché, il est nécessaire d'adapter le programme à la langue et la culture de chacun des groupes d'utilisateurs. Ceci n'est aucunement dû à un chauvinisme déplacé, mais répond au contraire à des exigences d'efficacité : l'utilisateur doit pouvoir se servir d'un programme dans sa langue de prédilection afin de ne pas rencontrer de problèmes de compréhension de l'interface et de pouvoir utiliser le programme de façon optimale.

La localisation comprend une grande part d'adaptation (d'après Schumann, 2000) :

- adaptation linguistique : outre le changement des composantes textuelles, il peut s'avérer nécessaire de procéder à des adaptations de formats numériques (signe décimal, point ou virgule), de format des dates (jj.mm.aaaa ou mm.jj.aaaa),
- adaptation audiovisuelle : il convient d'utiliser des couleurs, symboles et sons usuels et compréhensibles,
- adaptation technique : le logiciel doit être disponible dans un format qui permet de l'installer sur les systèmes les plus communs sur le marché ciblé,

- adaptation légale : le produit doit respecter la législation en vigueur dans le pays où il doit être commercialisé.

Le logiciel en question a été développé par Forgis ("Forschungsgemeinschaft für Integrierte Systemlösungen"), un institut allemand d'informatique appliquée, et s'adresse aux constructeurs automobiles européens. Nous avons été chargés de l'adaptation linguistique, soit la traduction des différents menus et boutons. Les textes apparaissant dans une interface graphique ne font pas partie intégrante du programme informatique : il serait fastidieux de chercher dans le code du programme les parties à traduire. Ce serait par ailleurs risquer de modifier par mégarde le programme. L'ensemble des composantes textuelles, dans toutes les langues, est enregistré dans des fichiers nommés "fichiers ressources" auxquels le programme renvoie par un système de références. Ce sont ces fichiers qu'il nous a été demandé de traduire de l'allemand vers le français.

Extrait de fichier "ressources"

```
//ToolBar
//*****
*****
1001, "Datenbank"
1002, "Speichern"
1003, "Drucken"
1004, "Suchen"
1005, "Modul"
1006, "Berichte"
1007, "Ende"
//ToolTips
1021, "Datenbank öffnen"
```

Toutes les entrées en allemand commencent par "1", celle en anglais par "2", "3" correspond au français et "4" à l'espagnol.

Ainsi, la ligne 1004, "Suchen" correspond à 3004, "Rechercher".

### 1.2.2.3 Utilisation d'un logiciel de TAO

Comme les fichiers ressources sont relativement répétitifs, il est judicieux d'utiliser un logiciel de TAO, ici "Déjà Vu", pour assurer la cohérence de la traduction. Le contenu des fichiers est exporté en format RTF et importé dans "Déjà Vu Interactive", un module du programme qui permet la gestion des projets de traduction.

"Déjà Vu" fonctionne selon le principe des mémoires de traduction : c'est une base de données de paires de phrases en langue source et cible accompagnées d'informations relatives au projet de traduction duquel les phrases sont issues.

Dans notre cas, la mémoire de traduction est vide. On commence par définir un projet de traduction dans "Déjà Vu Interactive" puis on y importe tous les fichiers à traduire. L'interface de "Déjà Vu Interactive" se compose d'un tableau à deux colonnes, l'une pour la langue source, l'autre pour la langue cible. On peut, au fur et à mesure de la traduction, alimenter la mémoire de traduction et propager la traduction au reste du fichier, c'est-à-dire que les phrases ou segments identiques en langue source sont retrouvés et leur traduction insérée dans les cellules correspondantes en langue cible.

Cette localisation a, pour certains termes très spécifiques, requis une grande part de recherches. Ce fut en particulier le cas pour les termes désignant certaines pièces mécaniques ou des matériaux utilisés dans l'industrie automobile. Pour lever les ambiguïtés, nous nous sommes référés à des sites Internet spécialisés.

- 3-2-1 Auto propose un glossaire de termes utiles à l'entretien des véhicules
- Motorlegend ("au service de la passion automobile") explique avec force détails et illustrations le fonctionnement des différents organes d'une voiture, allant même jusqu'à retracer l'évolution technique afin de faciliter la compréhension.

Une fois la traduction achevée, la colonne "langue cible" est exportée et intégrée aux fichiers ressources. Dans une interface graphique de logiciel, les noms des menus sont en principe relativement courts pour pouvoir être affichés dans l'espace qui leur est imparti. Nous n'avons pas respecté cet impératif de concision en conservant articles et prépositions afin de ne pas créer d'ambiguïtés artificielles. Ce sont les concepteurs du programme qui procéderont aux adaptations.

### **Note : Etat d'avancement de la proposition de directive**

Après plusieurs amendements proposés par le Parlement Européen et le Conseil, les deux organes sont parvenus à rédiger un projet commun le 23 mai 2000. Il est actuellement soumis au Parlement et au Conseil pour adoption finale.

## 1.3 Terminologie

### 1.3.1 Création d'une liste de termes multilingue

Dans le cadre des services proposés aux entreprises et de ses activités ayant trait à la terminologie, l'IAI a répondu à la demande du service de traduction du constructeur automobile Mercedes. Le service de traduction ne possède pas de base de données terminologiques, mais dispose de nombreux textes (techniques et publicitaires) traduits en plusieurs langues (allemand, anglais, français, espagnol). Il faudrait utiliser ces textes afin de créer une terminologie par un procédé semi-automatique. Le procédé est testé en premier lieu sur un seul texte pour l'allemand, l'anglais, le français et l'espagnol. S'il s'avère concluant, des textes supplémentaires seront traités.

#### 1.3.1.1 Extraction de termes

Il faut tout d'abord extraire les termes des textes de chaque langue. Pour cela, on utilise un script qui, à partir d'une analyse morphologique, écrit dans un fichier tous les syntagmes nominaux du texte accompagnés de leur contexte, c'est-à-dire de la phrase dans laquelle ils apparaissent. Le script "make\_term\_glossary" est très performant pour des textes en allemand, un peu moins pour les autres langues. Il génère entre autres les fichiers ListeTC\_DE et ListeTC\_FR, classés par ordre alphabétique. La vérification de la liste de termes est confiée à des "native speakers" de chaque langue.

#### 1.3.1.2 Appariement

Dans un deuxième temps, il faut faire correspondre à chaque terme allemand ses équivalents dans les autres langues et créer un fichier comprenant pour chaque entrée en allemand, son contexte, les termes étrangers équivalents et les contextes correspondants.

Nous sommes chargés d'établir la liste allemand-français.

Disposant de textes exactement parallèles dans les deux langues Texte\_DE et Texte\_FR, nous décidons de les aligner et de nous servir de l'alignement pour composer la liste de termes bilingue, comme le résume le schéma page suivante.

## Création d'une liste bilingue de termes à partir de textes alignés

Liste de termes et contextes allemands (générée automatiquement)

Liste de terme et contexte français préparée automatiquement, à compléter manuellement

German	French
Anstelle der Kugelumlauf-Lenkung des Fernlicht-Spotscheinwerfer	A la place de la direction à billes dont était équipé spot de feux de route
Fundstelle:	Fundstelle:
Hinter den Kunststoff-Streuscheiben der äußeren Schutzleisten	les diffuseurs en matière plastique des projecteurs baguette de protection
Fundstelle:	Fundstelle:
Die zusätzlichen Schutzleisten in den Eckbereichen	Les baguettes de protection supplémentaires
Wandlerüberbrückungs-Kupplung	embrayage avec pontage de convertisseur
Fundstelle:	Fundstelle:
Das bewährte Fünfgang-Automatikgetriebe mit Auspuffrohr	La boîte automatique à cinq sortie d'échappement
Fundstelle:	Fundstelle:

Anstelle der Kugelumlauf-Lenkung des Vorgängermodells erhält der neue CL die Zahnstangen-Lenkung der S-Klasse, die bei gleicher Präzision deutliche Gewichtsvorteile bietet.

A la place de la direction à billes dont était équipé l'ancien modèle, le nouveau CL reçoit la direction à crémaillère de la Classe S.

Recherche dans le texte du contexte allemand et de son correspondant français

German	French
27. September 1999	Le 27 septembre 1999
Der neue Mercedes-Benz CL	Le nouveau Mercedes-Benz CL
Inhalt	Sommaire
Seite	Page
Kurzfassung	En bref
Einzigartige Synthese von Fahrdynamik und Komfort	Alliance unique entre confort et dynamique de
Auf einen Blick	En un coup d'il
Technische Innovationen an Bord des CL	Innovations techniques à bord du CL
Auf ein Wort	En un mot
27. September 1999	Le 27 septembre 1999

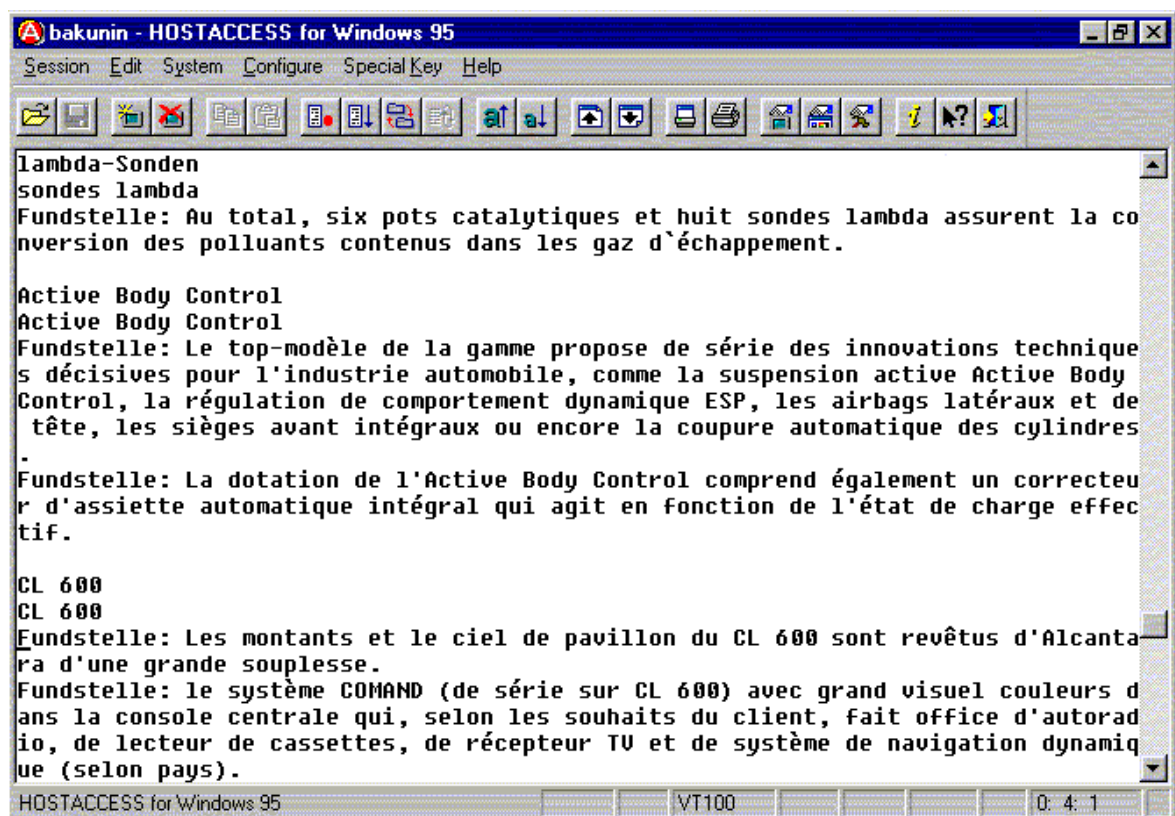
Textes allemand et français alignés

Nous utilisons le module d'alignement du logiciel "Déjà Vu", un équivalent du WinAlign de Trados Translator's Workbench. Il suffit d'importer les deux textes (Texte\_DE et Texte\_FR) dans le logiciel qui procède à l'alignement en reconnaissant les fins de phrases ou de propositions ainsi que différents marqueurs. Le programme se base pour cela sur des données statistiques (longueur moyenne d'un mot ou d'une phrase) ainsi que sur des composantes textuelles (noms propres, acronymes, chiffres).

Il faut ensuite vérifier que les textes correspondent bien et, au besoin, modifier la segmentation pour parfaire l'alignement. Les textes alignés segment par segment sont ensuite transférés dans la base de données (mémoire de traduction) de-fr.mdb. La liste des termes et contextes allemands ListeTC\_DE est ensuite importée dans Déjà Vu et "prétraduite" par le logiciel qui utilise la mémoire de traduction de-fr.mdb construite précédemment. La plupart des contextes sont trouvés par le programme, il suffit de chercher les autres dans la base de données comme l'illustre le schéma ci-dessous.

Une fois la liste en français complétée, il suffit d'exporter les deux colonnes se correspondant parfaitement en format ASCII, puis quelques manipulations sous Unix (voir section Unix) permettent de créer un fichier prêt à être importé dans un logiciel de base de données terminologiques du type Trados MultiTerm.

Liste définitive allemand-français



Cette base de données créée sous MultiTerm sera présentée à Mercedes. Si elle répond aux besoins du service de traduction, il sera envisagé d'automatiser le processus effectué afin de l'appliquer à un grand nombre de termes pour créer une base de données terminologique globale pour Mercedes.

Ces travaux ont été effectués sur des documents de formats différents (texte vi, pages html, base de données de Déjà Vu), ce qui nous a permis de nous familiariser avec les divers outils permettant de les traiter.

Il nous a également été demandé d'améliorer la grammaire française de CAT2, système de traduction automatique développé dans le cadre du projet EUROTRA. Cette grammaire est utilisée pour l'analyse des phrases lorsque le français est langue source, pour la génération de la traduction quand il s'agit de la langue cible. Les améliorations apportées par Patrick Da Costa concernent en particulier la génération des formes fléchies. De nouvelles règles permettent la formation du féminin et du pluriel des adjectifs, du pluriel des substantifs et de toutes les personnes dans la conjugaison des verbes au présent.

Modifier une grammaire d'un système de traduction est un exercice exigeant une grande rigueur et beaucoup de réflexion : un changement d'apparence minime peut se révéler important dans une phase ultérieure. Il est très formateur d'élaborer soi-même de nouvelles règles et d'observer comment elles modifient le comportement du système de traduction, en améliorant son efficacité ou au contraire en la diminuant, et de persévérer jusqu'à ce que les progrès soient sensibles.

## 2 Diagnostic d'une base de données terminologiques existante

Le 30 mai 2000, un contrat a été passé entre l'IAI et le service de traduction de Boehringer Ingelheim (industrie pharmaceutique).

Il est convenu que l'IAI effectue un diagnostic de la base de données terminologiques de Boehringer Ingelheim et livre les résultats de l'analyse à la fin du mois de juin sous la forme d'un rapport.

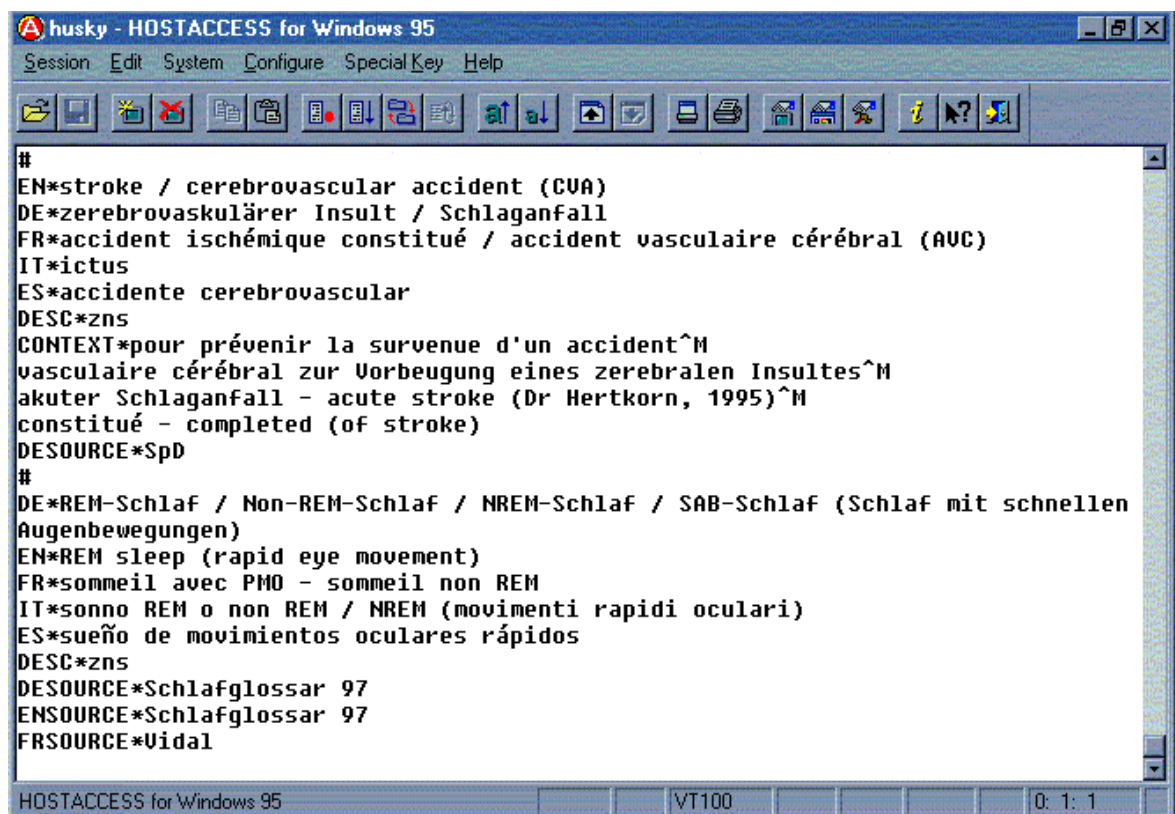
J'ai participé aux phases d'analyse des données et de présentation des résultats sous la direction du Docteur Axel Theofilidis.

### 2.1 Situation de départ

#### 2.1.1 Données de Boehringer Ingelheim

Le service de traduction de Boehringer Ingelheim stocke ses données terminologiques à l'aide du logiciel Termbase. Afin de pouvoir traiter les données sous environnement de travail Unix, les fiches sont exportées de Termbase en format ASCII. On extrait les noms des champs ainsi que leur valeur et on les enregistre dans un fichier nommé DB\_ori.

Exemple de fiche terminologique exportée :



```
husky - HOSTACCESS for Windows 95
Session Edit System Configure Special Key Help
#
EN*stroke / cerebrovascular accident (CUA)
DE*zerebrovaskulärer Insult / Schlaganfall
FR*accident ischémique constitué / accident vasculaire cérébral (AVC)
IT*ictus
ES*accidente cerebrovascular
DESC*zns
CONTEXT*pour prévenir la survenue d'un accident^M
vasculaire cérébral zur Vorbeugung eines zerebralen Insultes^M
akuter Schlaganfall - acute stroke (Dr Hertkorn, 1995)^M
constitué - completed (of stroke)
DESOURCE*SpD
#
DE*REM-Schlaf / Non-REM-Schlaf / NREM-Schlaf / SAB-Schlaf (Schlaf mit schnellen
Augenbewegungen)
EN*REM sleep (rapid eye movement)
FR*sommeil avec PMO - sommeil non REM
IT*sonno REM o non REM / NREM (movimenti rapidi oculari)
ES*sueño de movimientos oculares rápidos
DESC*zns
DESOURCE*Schlafglossar 97
ENSOURCE*Schlafglossar 97
FRSOURCE*Uidal
HOSTACCESS for Windows 95 VT100 0: 1: 1
```

Dans le cadre du contrat passé, seules les données en allemand (codées DE\*) vont être analysées. Les exemples sont cités ici uniquement en allemand car ils sont tirés de la

base de données de Boehringer Ingelheim et illustrent des phénomènes propres à cette langue.

### 2.1.2 Ressources de l'IAI

Des analyses de bases de données terminologiques ont déjà été effectuées pour d'autres entreprises par le Docteur Axel Theofilidis. Les premiers travaux ont été menés en utilisant des outils dont l'IAI disposait déjà : MULTIDOC contient des modules de vérification d'orthographe, de grammaire, de style et de synonymes. Peu à peu, d'autres ressources informatiques sont développées spécialement pour les travaux de terminologie. La conversion des données de départ en fichier ASCII permet leur traitement à l'aide de commandes et de programmes Unix.

Le principe de l'analyse est schématisé page suivante par le diagramme :

"Analyse de la base de données terminologique de Boehringer Ingelheim  
**Diagramme de principe**".

## 2.2 Analyse préalable de la base de données

Le fichier de départ sous Termbase est une base de données dont toutes les fiches n'ont pas exactement la même structure. On peut y trouver les champs suivants :

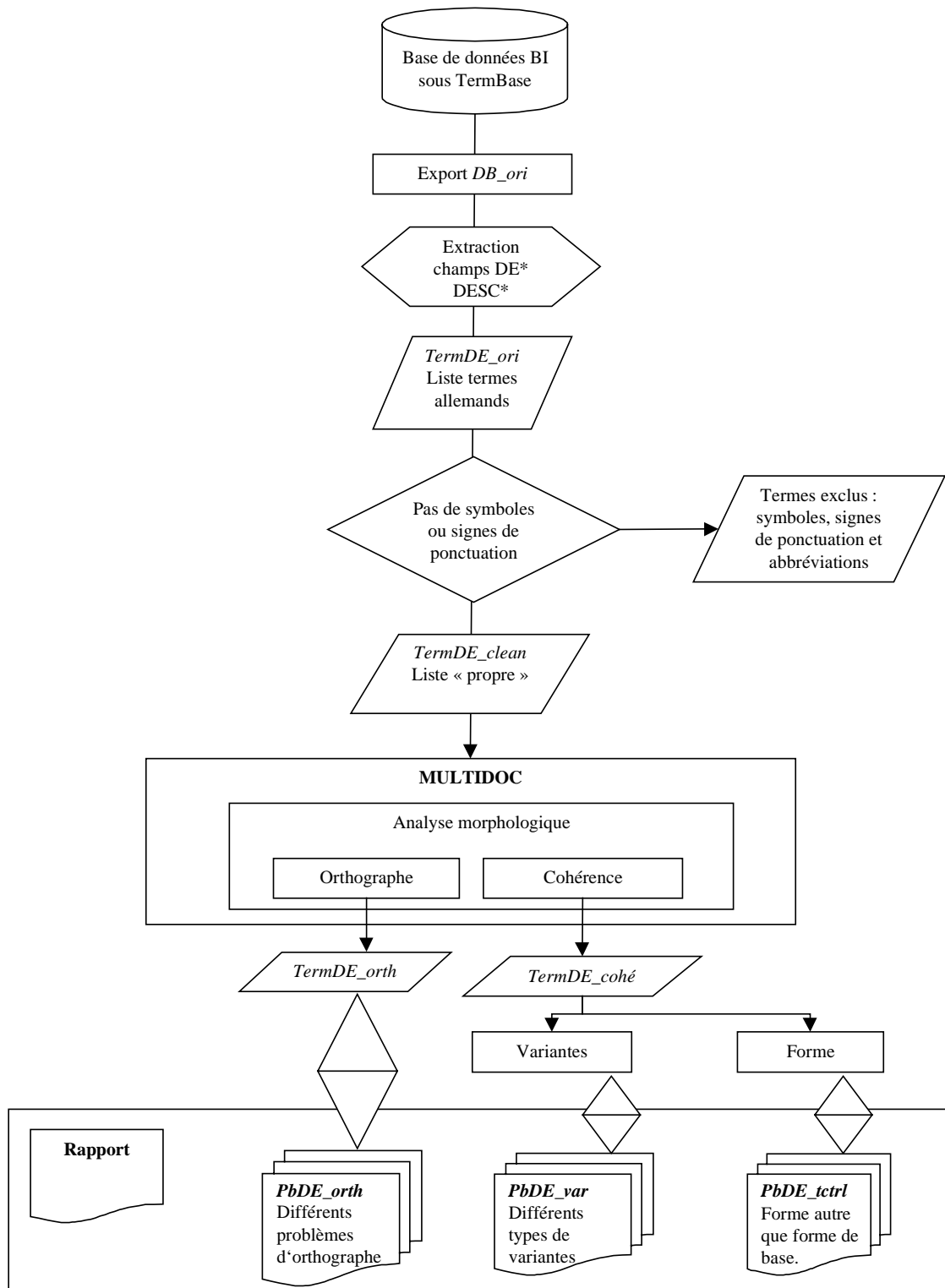
Code du champ	Contenu
DE*	Terme allemand
EN*	Terme anglais
ES*	Terme espagnol
FR*	Terme français
IT*	Terme italien
DESOURCE*	Code de la source du terme allemand
ENSOURCE*	Code de la source du terme anglais
ESSOURCE*	Code de la source du terme espagnol
FRSOURCE*	Code de la source du terme français
ITSOURCE*	Code de la source du terme italien
DESC*	Descripteur : information sur le terme <sup>°</sup>
CONTEXT*	Contexte : exemple de l'utilisation du terme <sup>°</sup>

<sup>°</sup> Champ non dépendant de la langue

Chaque champ ne peut apparaître qu'une seule fois par fiche. Il semble cependant qu'aucune fiche ne comporte la totalité de ces champs accompagnés d'une valeur correspondante.

Du fichier original ont été extraites toutes les entrées en allemand accompagnées de leur descripteur (champs DE\* et DESC\*) et enregistrées dans le fichier TermDE\_ori. Ces données à analyser représentent 48 642 entrées.

## Analyse de la base de données terminologique de Boehringer Ingelheim *Diagramme de principe<sup>1</sup>*



<sup>1</sup> Voir légende en Annexe A

La base de données est passée en revue afin de détecter les problèmes les plus flagrants.

On remarque fréquemment que plusieurs variantes sont codées dans un même champ et séparées par différents signes de ponctuation (virgule, point virgule, trait d'union, slash). D'autre part, on trouve des entrées présentant des symboles tels que parenthèses, crochets, étoile, etc. On note également que de nombreuses entrées codées "DESC\*abb" (pour abréviation) ne sont pas des abréviations.

On procède à une sorte de filtrage en utilisant un script, une suite de commande Unix qui forment un petit programme.

Le script "do\_tlist" permet d'exclure d'un fichier les lignes présentant une propriété déterminée. Nous décidons ici de trier toutes les lignes contenant le descripteur "abb" (abréviation), ainsi que celles contenant des séparateurs (; - /) ou des caractères spéciaux (.:?!\*()[]{}<>"). En effet, dans l'étape suivante, l'analyse automatique ne pourrait donner de résultats cohérents dans de tels cas.

Il reste alors 39 346 entrées dans le fichier TermDE\_clean à soumettre à l'analyse morphologique.

## 2.3 Analyse des termes

La liste des termes restants, TermDE\_clean, est envoyée à MULTIDOC pour analyse.

L'analyseur morphologique MPRO est une des composantes du programme MULTIDOC. Il effectue les opérations suivantes :

- décomposition en mots,
- analyse morphologique,
- désambiguïsation des homographes.

Avec un fichier analysé, on obtient deux fichiers de sortie : le premier comprend les problèmes ayant trait à l'orthographe (TermDE\_orth), le second répertorie différents types de variantes (TermDE\_cohé).

### 2.3.1 Orthographe

Les termes de la liste de départ sont soumis à un contrôle de l'orthographe. Ceci permet d'une part de reconnaître tous les cas relevant de la réforme de l'orthographe allemande (ß et ss, ph et f, etc.) et d'autre part, de répertorier tous les termes pour lesquels l'analyse linguistique échoue. C'est le cas lorsque le terme ou un de ses éléments n'est pas reconnu par le dictionnaire auquel fait appel l'analyseur morphologique : il peut s'agir d'une erreur de frappe, d'un mot étranger, d'une marque déposée, d'un nom propre ou d'une abréviation.

Le fichier de sortie TermDE\_orth comprend tous les termes de la liste de départ (TermDE\_clean) ainsi que des balises codant les problèmes d'orthographe.

## 2.3.2 Cohérence interne

MULTIDOC dispose également d'un vérificateur de cohérence interne du texte. Celui-ci doit signaler au rédacteur qu'il a employé des termes semblables mais non identiques, et demander une confirmation : s'agit-il bien de termes différents se rapportant à des notions différentes ?

Ce programme permet de retrouver dans la base de données différents types de variantes par un processus interne de comparaison. Il effectue une analyse d'une part sur les variantes possibles des termes pour trouver les différentes dénominations d'une même notion, et d'autre part sur la forme des termes. Les résultats sont enregistrés dans le fichier TermDE\_cohé.

### 2.3.2.1 Types de variantes

Un premier groupe de données comprend ainsi les variantes **morpho-graphémiques** :

- variantes de composition : Abstandhalter - Abstandshalter ,
- variantes orthographiques : Äthylchlorid - Ethylchlorid ,
- variantes de dérivation : Ablösen - Ablösung .

Si on indique au module de vérification un fichier de synonymes pour le domaine en question, il fournit pour la base de données une liste de **synonymes possibles**. Le fichier de synonymes peut être défini en fonction du profil de l'utilisateur (et du domaine), ainsi, si on a défini la paire de synonymes Anlage - Vorrichtung, le programme déduit que Abfüllanlage est un synonyme de Abfüllvorrichtung.

De plus, le vérificateur indique les **variantes de réduction** des termes complexes, c'est-à-dire un terme composé d'un certain nombre d'éléments et un autre très semblable, comportant moins d'éléments. Ainsi Fluoreszenzphotometer serait une variante de réduction possible de Fluoreszenzspektralphotometer, selon le modèle ABCD -> A\_CD; de même pour Aluminiumfolie et Aluminiumverbundfolie selon le modèle ABC -> A\_C.

### 2.3.2.2 Forme des termes

Dans une base de données terminologiques, les termes doivent être entrés sous leur forme de base ou forme lexicographique : verbes à l'infinitif et substantifs au nominatif singulier. Ce n'est pourtant pas toujours le cas dans la pratique.

Un module supplémentaire effectue donc sur la liste un contrôle de la forme de termes et détecte ainsi tous les termes qui sont présents dans la base de données sous une forme autre que leur forme lexicographique. Ce sont :

- des substantifs commençant par une minuscule au lieu d'une majuscule,
- des verbes ou adjectifs commençant par une majuscule,
- des substantifs présentant une flexion (marque de cas ou de pluriel),
- et dans une moindre mesure, des infinitifs nominalisés et des verbes dans une forme finie.

Comme le programme calcule pour chacun de ces mots sa forme de base, on la compare au fichier original de la base de données afin de signaler les termes apparaissant deux fois, sous une forme erronée et sous leur forme de base, ainsi que ceux qui n'apparaissent que sous une forme erronée.

Un autre module permet de vérifier l'utilisation de la règle d'utilisation des traits d'union dans les mots composés en allemand. La nouvelle orthographe allemande stipule par exemple que les mots composés de trois termes ou plus peuvent s'écrire avec un trait d'union afin de faciliter la lecture et de limiter les ambiguïtés : Blutzuckerbelastungskurve pourrait ainsi également s'écrire Blutzucker-Belastungskurve.

On met ensuite en parallèle les résultats proposés par l'analyseur morphologique et les données présentes dans la base.

## 2.4 Résultats

Chaque fichier fourni par les programmes de vérification est ensuite mis en ordre et en forme. De plus, le rapport décrit le processus d'analyse et le contenu des fichiers résultats. Ces données définitives sont envoyées au client par courrier électronique sous forme d'archive zip.

L'utilisation de quelques filtres (commandes 'grep' et 'sort', cf. Manipulation de données en utilisant des filtres) permet de classer les problèmes en différentes catégories.

Les problèmes d'orthographe du fichier PbDe\_orth sont classés en quatre groupes, eux-mêmes divisés en sous-groupes :

- cas dépendants de la réforme (nouvelle/ancienne orthographe),
- problèmes de casse (majuscule/minuscule),
- mots inconnus (marques, noms propres, erreurs de frappe etc.),
- entrées présentant plusieurs de ces problèmes.

Les variantes possibles de termes sont regroupées dans le fichier PbDE\_var et classées en :

- variantes morfo-graphémiques,
- synonymes possibles,
- variantes de réduction de termes composés.

Les termes pour lesquels l'analyse révèle un problème de forme se retrouvent dans le fichier PbDE\_tctrl :

- présence d'une marque de cas,
- présence d'une marque de pluriel,
- termes qui devraient commencer par une majuscule,
- mots composés de quatre radicaux ou plus sans trait d'union.

Les résultats sont livrés à Boehringer Ingelheim sous la forme d'un rapport détaillant les analyses effectuées et les types de problèmes rencontrés dans la base de données, accompagnés d'une analyse statistique.

Au cours de l'analyse, il nous a été donné de constater que de nombreux problèmes de cohérence sont dus à un manque de concertation au sein même du groupe des utilisateurs de la base de données. L'élaboration de quelques règles simples permettrait, malgré la structure linéaire du logiciel Termbase, une utilisation bien plus efficace de la base de données. La conclusion du rapport donne des pistes dans ce sens.

**Pour la base de données dans son ensemble :**

- entrer uniquement un terme par champ-vedette, n'utiliser ni symboles, ni signes de ponctuation,
- entrer systématiquement les abréviations, de préférence dans un champ séparé,
- entrer systématiquement synonymes et variantes, de préférence dans un champ séparé.

**Pour l'allemand :**

- écrire les "Umlaute" (ü et non ue),
- entrer les termes sous leur forme lexicographique (substantifs au nominatif singulier),
- écrire les termes relevant de la langue courante selon la nouvelle orthographe en vigueur (la réforme de l'orthographe ne s'applique qu'à la langue courante et n'a pas l'ambition de réglementer les langues de spécialité).

Les données font l'objet d'une dernière analyse, à visée interne uniquement. Il s'agit d'une analyse statistique de la composition des termes complexes, noms composés et expressions de plusieurs mots. La nature de chaque élément du terme est reconnue, on obtient ainsi des modèles ou schémas de formation des termes. On cherche ainsi à observer les phénomènes de composition afin de pouvoir en déterminer la systématique et d'arriver, dans un futur plus ou moins proche, à une "réglementation des néologismes".

### 3 Création d'une base de données terminologiques

Le besoin de réutiliser une terminologie précise se rapportant à un domaine donné augmente en même temps que les progrès technologiques. Devant l'omniprésence de l'informatique, l'utilisation du format électronique pour le stockage des données terminologiques s'est imposé depuis longtemps.

Le besoin de normalisation s'est ensuite très vite fait sentir. Si les premières normes consacrées à la terminologie datent des années quatre-vingt, celles portant sur les "aides informatiques en terminologie" ne sont parues qu'en 1997. Le comité technique 37 de l'Organisation Internationale de Normalisation (ISO/TC 37) est spécialisé dans les questions de terminologie.

La terminologie est également une préoccupation dans l'industrie, où les documents et notices techniques doivent être d'excellente qualité pour des raisons de sécurité et de responsabilité légale. C'est dans cette optique que l'IAI développe, en collaboration avec des partenaires industriels, un logiciel de gestion terminologique. Celui-ci doit permettre de vérifier l'utilisation cohérente des termes dans les documents élaborés par le rédacteur technique en se basant sur la base de données de l'entreprise. Une autre fonctionnalité assistera le processus de validation des termes, depuis la proposition d'un néologisme par un rédacteur technique jusqu'à son acceptation par le service de terminologie et son entrée en tant que terme dans la base de données de l'entreprise.

Il existe aujourd'hui un grand nombre d'outils informatiques destinés à la gestion de bases de données terminologiques. Citons MultiTerm (Trados) et TermStar (Star). S'ils possèdent un grand nombre de paramètres prédéfinis, ils ne suffisent pas à eux seuls à concevoir une base de données terminologiques correcte.

L'objectif d'une base de données terminologiques est de stocker des informations terminologiques, afin de les rechercher et surtout de pouvoir les retrouver facilement.

La norme ISO 1087:1990 donne la définition suivante :

**Base de données terminologiques :**

ensemble structuré de données terminologiques constitué en système d'information électronique.

Toute la subtilité se cache dans le terme "structuré". Les données sont en effet structurées à deux niveaux :

- au niveau de chaque notion, par un système de fiche,
- au niveau de la base de données, grâce à un système de relations entre les fiches.

Il importe avant tout de bien définir ces structures qui constituent le "squelette" de la base de données. Les erreurs et incorrections du départ sont difficilement modifiables dans des phases ultérieures du travail terminologique.

## 3.1 Structure d'une fiche terminologique

### 3.1.1 Définitions

Pour reprendre les plus anciennes techniques de stockage de terminologie (fiches bristol classées par ordre alphabétique dans une boîte à chaussure), on utilise également dans les terminologies électroniques un système de fiches.

Reprenons ici les définitions données par la norme ISO 1087:1990 afin de clarifier les choses avant de poursuivre ce propos.

**Notion :**

Unité de pensée constituée par abstraction à partir des propriétés communes à un ensemble d'objets.

*Note – Les notions ne sont pas liées aux langues individuelles. Elles sont cependant influencées par le contexte socioculturel.*

**Terme :**

Désignation au moyen d'une unité linguistique d'une notion définie dans une langue de spécialité.

*Note – Un terme peut être composé d'un ou plusieurs mots (terme simple ou terme complexe) et même de symboles.*

Ces définitions sont communément admises et utilisées lorsqu'il est question de travail terminologique.

La plupart des théories terminographiques préconisent un système notionnel, soit une fiche par notion, contrairement à la lexicographie qui travaille par mot. Dans un système terminographique, il existe une fiche par terme.

Les homonymes, se rapportant à des notions différentes, seront donc répertoriés dans des fiches différentes dans un système terminographique mais dans une seule entrée dans un système lexicographique.

### 3.1.2 Catégories de données

Une fiche terminologique comprend

- des données de gestion de la base : date de création de la fiche, numéro de la notion, date de modification, auteur de la modification, etc.,
- des données se rapportant à la notion : domaine, sous-domaine,
- les données terminologiques à proprement parler : terme, synonymes...

Précisons à présent la nature des données terminologiques. Ce sont des informations de différents types, codées dans les fiches sous forme de champs.

**Désignation**

C'est le champ central de la fiche terminologique, celui qui comprend le terme privilégié désignant la notion en question. Ce terme peut se composer d'un ou plusieurs mots, il peut donc être simple ou complexe. Il est entendu que les termes doivent être répertoriés sous leur forme de base (en général le singulier pour les substantifs, l'infinitif pour les verbes).

### **Informations grammaticales**

Ce champ contient des indications sur la partie du discours, le genre et le nombre du terme privilégié. Ces informations contribuent à la désambiguïsation des homographes (en particulier en anglais où les formes sont très ambiguës). Une indication de la formation du pluriel du terme peut être envisagée afin de clarifier le pluriel des mots composés et de permettre à des utilisateurs d'une autre langue maternelle une utilisation optimale de la base de données.

### **Définition**

Description et explication de la notion.

### **Contexte**

On rapporte ici le terme dans un usage authentique. Ce champ a pour but de préciser entre autres les collocations dans lesquelles apparaît le terme. Il est bien entendu indispensable d'utiliser comme contexte des données réelles, tirées de sources sûres : l'auteur doit avoir écrit dans sa langue maternelle et sur un sujet qui lui est familier.

### **Synonymes**

On entre ici les autres désignations acceptées pour la notion. Les synonymes peuvent jouir du même degré d'acceptation que le terme privilégié, d'un degré d'acceptation moindre (ce sont alors des termes tolérés) ou peuvent être à éviter (termes à rejeter). Il est de plus recommandé d'entrer également chaque synonyme comme terme dans une fiche séparée, avec éventuellement un renvoi à la fiche du terme privilégié afin de faciliter les recherches.

### **Source**

Ce champ doit accompagner chacune des données précédemment citées. Les sources peuvent être indiquées sous forme codée.

### **Code du terme**

Il peut être formé à partir du numéro du projet ou du contrat, et de celui de la notion dans le système. Ce code permet d'identifier chaque terme de façon unique dans la base de données.

Une fiche peut en outre comprendre, selon les besoins spécifiques de l'utilisateur, des champs tels que :

- variantes régionales du terme (par exemple dans les communautés francophones de Suisse ou de Belgique),
- illustrations (accompagnées de leur source),
- remarques : toutes les indications utiles à l'utilisateur et n'apparaissant pas dans les autres champs (indication de style, formules, commentaires sur la notion n'apparaissant pas dans la définition citée, etc.)

Ces données sont à comprendre dans le cas d'une fiche terminologique monolingue. Dans une base de données multilingue, il faudrait reprendre une structure identique pour chacune des langues, en ajoutant pour chaque terme une indication de la langue concernée. En effet, une base de données multilingue idéale ne privilégie aucune des langues, afin de permettre de travailler dans toutes les directions. Dans les logiciels de gestion de bases de données (type MultiTerm), il est possible de permuter langue source et langue(s) cible(s) de la base.

La récente norme ISO 12620:1999 régit le format des catégories de données de façon exhaustive.

### 3.2 De la base de données au système notionnel

Une fiche terminologique du type décrit ci-dessus peut être utilisée quelle que soit l'intention de son concepteur : traducteur pour des besoins de traduction ou terminologue effectuant un travail systématique. Dans ces deux cas, les procédés de travail sont en effet radicalement différents, comme l'illustre le tableau ci-dessous (d'après Wright, 1997).

Travail terminologique ponctuel (traducteur)	Travail terminologique systématique (terminologue)
<ul style="list-style-type: none"> <li>• extraction des termes de textes isolés</li> <li>• création de fiches à partir des notions présentes</li> <li>• indication des contextes</li> <li>• si le temps le permet, reconstruction d'un système de notions sur les bases présentes</li> </ul>	<ul style="list-style-type: none"> <li>• collection la plus exhaustive possible de termes sur un domaine dans son ensemble</li> <li>• construction d'un système de notions</li> <li>• recherche de définitions pertinentes</li> <li>• liaison des fiches dans un système notionnel</li> </ul>

Le traducteur, s'il doit répondre à un besoin immédiat, peut utiliser les fiches d'une base de données pour consigner le résultat de ses recherches, tout en admettant que certains champs restent vides. Ils pourront être remplis au cours de recherches ultérieures, lorsque la charge de travail réduite permet d'effectuer des travaux complémentaires.

Le terminologue, lui, cherche à avoir une vue d'ensemble d'un domaine. Il lui faut donc tout d'abord définir avec précision ce domaine avant de procéder à la phase de collection des données. Il lui faut ensuite établir les liens reliant ces notions afin de pouvoir construire son système notionnel. Dans un souci d'exactitude, il peut faire appel à un spécialiste du domaine.

#### **Relations entre notions**

Les liens entre les notions ou relations notionnelles sont difficilement répertoriables de façon exhaustive. On distingue cependant deux grandes familles de relations :

- relations hiérarchiques, qui existent entre deux niveaux d'un système notionnel (entre une notion super-ordonnée et ses notions subordonnées). Les relations "partie-tout" et "espèce-genre" sont deux exemples de relations hiérarchiques.
- relations coordonnées, au même niveau d'un système notionnel. On distingue, entre autres, relations séquentielles et relations pragmatiques. Les relations séquentielles existent entre "objets présentant une contiguïté spatiale ou temporelle", comme par exemple les relations "cause-effet", "étapes d'un processus" ou "producteur-produit" (ISO 1087:1990).

Relations hiérarchiques et coordonnées ont été plus précisément définies dans les normes ISO 704:1987 et DIN 2330:1979.

Il existe également à l'intérieur des termes des relations fonctionnelles qu'il serait intéressant d'étudier afin de rédiger une typologie des relations entre les différents composants d'un terme complexe.

S'il est, dans la pratique d'une base de données informatique, difficile (mais non impossible) de mettre en œuvre une telle diversité de liens, l'hypertexte permet toutefois de créer des renvois logiques entre fiches et de s'approcher ainsi du système notionnel.

### 3.3 Considérations d'ordre pratique

Ces pistes sur la création d'une base de données terminologiques ne seraient pas complètes sans quelques considérations pratiques, fruit de l'observation de bases de données existantes.

#### 3.3.1 Nature des champs

Dans une fiche terminologique électronique, on distingue différents types de champs.

##### **Champs d'index**

Ils sont indexés, c'est-à-dire que c'est sur eux que vont s'effectuer les recherches ultérieures. Les champs "Désignation" et "Code du terme" sont des champs d'index.

##### **Champs de texte**

Comme leur nom l'indique, on peut y saisir un texte libre. Les champs "Définition" et "Contexte" sont des champs de texte.

##### **Champs d'attributs**

Ils ne peuvent contenir que des valeurs prédéfinies. Par exemple, le champ "Genre" ne peut contenir qu'une des valeurs suivantes : "masculin", "féminin" ou "neutre". Dans l'interface graphique, ces valeurs font partie d'un menu déroulant dans lequel il n'est possible de sélectionner qu'une seule ligne. Ce format de menu déroulant empêche une entrée manuelle des données qui pourrait être erronée : la valeur "masculin" pourrait être codée "masc." par un terminologue et "m." par un autre.

#### 3.3.2 Entrée des données dans les champs

##### **Forme des termes**

Dans une base de données terminologiques, il convient de coder le terme sous sa forme de base, encore appelée forme canonique ou forme lemmatisée. Ceci revient le plus souvent à choisir le singulier pour les substantifs et l'infinitif pour les verbes, à quelques exceptions près. En effet, comment entrer des termes pour lesquels l'usage privilégie le pluriel ou des expressions idiomatiques dans lesquelles le verbe apparaît à une forme finie ? Si le pluriel renvoie à une notion clairement différente du singulier, il est recommandé de considérer le pluriel comme un terme et de le traiter comme tel dans une fiche individuelle. Ce type de cas est à étudier en fonction de la pertinence et de la précision des recherches que le logiciel est à même d'effectuer.

D'autre part, il est, en français, conseillé de n'utiliser les majuscules que pour les noms propres, et de ne pas chercher à répondre à des critères d'esthétique en utilisant les majuscules au début de chaque terme, voire de chaque mot.

##### **Ordre des mots**

Contrairement aux dictionnaires qui font apparaître le mot "de tête" en début de chaîne afin de faciliter la recherche, les termes complexes peuvent être entrés dans une base de données terminologiques en suivant l'ordre naturel de leurs composants. En effet, les modules de recherche permettent le plus souvent d'effectuer des recherches faisant appel à des caractères génériques et il n'est alors pas nécessaire d'effectuer des permutations.

##### **Respect de l'accentuation**

En fonction des langues couvertes par la base de données, différents caractères peuvent apparaître. Ceux-ci reflètent les spécificités de chaque langue, il convient donc de les respecter absolument. Il faut tout d'abord que le logiciel soit adapté à un grand nombre de caractères. D'autre part, les utilisateurs sont invités à entrer des termes en respectant l'accentuation, ce qui implique, en fonction du type de clavier utilisé, d'avoir recours à des

moyens détournés (codes ASCII ou insertion de caractères spéciaux). Ceci permet de garantir l'exactitude de la base de données.

Ces considérations d'ordre pratique, ainsi que toutes les autres recommandations utiles pour l'établissement et la gestion d'une base de données terminologiques doivent faire l'objet d'une réflexion préalable chez le ou les utilisateurs futurs. En effet, si plusieurs utilisateurs effectuent des entrées et des modifications dans une base sans suivre une procédure standardisée, une systématisation réfléchie et approuvée, les données prendront rapidement un format des plus anarchiques et la base perdra sa raison d'être.

Le but d'une base de donnée est bien de retrouver plus rapidement des informations terminologiques, et non d'augmenter la durée des recherches.

Ces réflexions d'ordre général doivent ensuite faire l'objet d'une adaptation aux besoins de l'utilisateur. Selon le but visé (base multilingue stockant des traductions ou système notionnel), la structure et les règles méthodologiques changent sensiblement, des champs peuvent être ajoutés ou enlevés.

### 3.3.3 Un exemple : la base de données terminologiques du Service de la langue française du Ministère de la Communauté française de Belgique

La Commission "Terminologie" du Service de la langue française s'est donné pour but de déterminer dans quelle mesure la "terminologie produite en France correspond aux usages qui sont observés dans les milieux professionnels de la Communauté française de Belgique".

Pour cela, la Commission évalue l'implantation en Belgique des termes validés par la Délégation Générale à la Langue Française (DGLF). Ces termes, publiés au Journal Officiel, sont présents dans la base de données de la DGLF consultable en ligne.

Les règles régissant la structure de la base de la DGLF ont été adaptées pour répondre aux besoins du Service de la langue française. Citons quelques exemples.

Le champ **vedette** reprend le plus souvent le terme de l'arrêté français, sauf s'il bénéficie d'une mauvaise implantation en Communauté française de Belgique. Il ne contient en principe ni anglicisme, ni abréviation.

Le champ **synonyme** se veut le reflet des usages : il peut contenir des termes déconseillés par les arrêtés s'ils bénéficient d'une très bonne implantation en France et en Belgique : acronymes, sigles, voire exceptionnellement anglicismes à la condition que leur morphologie les rende intégrables en français.

The screenshot shows a Netscape browser window displaying the Terminobanque website. The browser's address bar shows the URL <http://www.cfwb.be/franca/bd/bd.htm>. The page title is "Terminobanque - Netscape". The website header includes "Service de la langue française" and "Banque de données terminologique" with a red rooster logo. A left-hand navigation menu lists various categories such as "Accueil", "Mode d'emploi", "Index", "Français", "Anglais", "Néerlandais", "Allemand", "Environnement", "Équipement", "Génie génétique", "Informatique", "Mer", "Nucléaire", "Personnes âgées", "Pétrole", and "Santé et médecine". The main content area displays the entry for "Toile d'araignée mondiale". The entry includes a list of related terms (e.g., "Toile mondiale", "Toile", "Tolérance aux pannes\*"), a domain ("Informatique/Internet"), a grammatical category ("n.f."), and synonyms ("Toile mondiale", "Toile", "T.A.M."). The definition states: "Dans l'internet, système, réparti géographiquement et structurellement, de publication et de consultation de documents faisant appel aux techniques de l'hypertexte." The source is cited as "J.O. du 16/03/99". A note explains the acronym T.A.M. and the term's usage. The entry also provides translations for "Anglais", "Néerlandais", and "Allemand". The browser's status bar at the bottom shows the date "Sonntag, 3. September 2000" and the time "19:12".

Cette base de données est un relativement bon exemple. Elle a été conçue par le Centre Termisti, le centre de recherche en terminologie de l'Institut supérieur de traducteurs et interprètes de Bruxelles, ce qui est déjà en soit un gage de rigueur et de méthodologie. Une structure précise a été définie suivant les objectifs du Service de la langue française et les fiches suivent effectivement cette systématique. Entre autres, l'utilisation de liens hypertexte permet de renvoyer à des notions voisines et de s'approcher ainsi de la structure d'un système notionnel tel que le conçoit Van Campenhoudt dans ses cours de terminologie.

## 4 Unix

J'aimerais résumer ici quelques-unes des notions et commandes que j'ai pu apprendre et utiliser en effectuant les travaux décrits dans les sections précédentes de ce rapport. Cette liste n'a aucunement la prétention d'être exhaustive, ce n'est qu'un petit aperçu des immenses possibilités d'Unix. Pour de plus amples informations, se référer au manuel en ligne d'Unix.

### 4.1 Notions de base

La structure d'Unix est basée sur un système de classement des fichiers en arborescence : chaque répertoire peut contenir des fichiers et sous-répertoires et ainsi de suite.

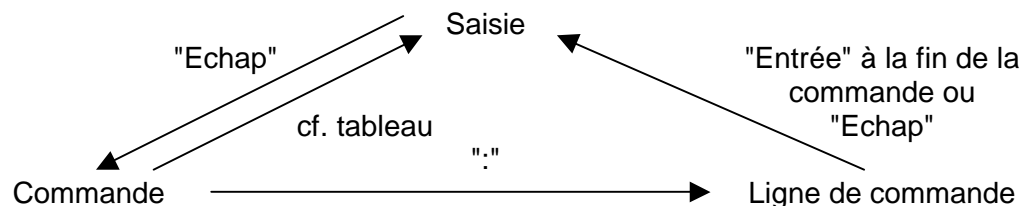
Les commandes les plus utiles lorsqu'on travaille avec des fichiers sont résumées ci-dessous.

Syntaxe	Explication
<code>cd répertoire</code>	va dans le répertoire indiqué ("change directory"). . est le répertoire courant, .. est le répertoire parent dans l'arborescence.
<code>ls</code>	liste le contenu d'un répertoire. L'option <code>-l</code> permet un affichage plus complet du contenu du répertoire, indiquant entre autre les droits et la taille du fichier, son créateur et sa date de dernière modification ("list").
<code>pwd</code>	indique le nom du répertoire courant ("print working directory").
<code>mkdir répertoire</code>	crée le répertoire indiqué ("make directory").
<code>mv fichier rép</code>	déplace le fichier dans le répertoire indiqué("move"). sert également à renommer les fichiers.

### 4.2 Commandes vi

vi est un éditeur de texte présent sous Unix. Il fonctionne en deux modes : saisie et commande. On utilise le clavier pour entrer le texte en mode saisie, et pour le modifier lorsqu'on est en mode commande ou sur la ligne de commande.

#### 4.2.1 Passage d'un mode à l'autre



### Passage de mode commande en mode saisie

Commande	Texte inséré...
a	après le curseur ("append")
i	avant le curseur ("insert")
A	à la fin de la ligne courante
I	au début de la ligne courante
o	sur une nouvelle ligne suivant la ligne courante
O	sur une nouvelle ligne précédent la ligne courante

### 4.2.2 Déplacement du curseur

Commande	Déplacement
<b>Caractère par caractère</b>	
h	gauche
j	bas
k	haut
l	droite
<b>Mot par mot</b>	
e	fin du mot courant ("end")
b	début du mot courant ("begin")
w	début du mot suivant ("word")
E	idem mais considèrent uniquement les espaces comme délimiteurs de mot (ni ponctuation, ni chiffre).
B	
W	
<b>Dans l'écran</b>	
Ctrl + d	un demi écran vers le bas ("down")
Ctrl + u	un demi écran vers le haut ("up")
Ctrl + f	un écran vers le bas ("forward")
Ctrl + b	un écran vers le haut ("backward")
<b>Dans le fichier</b>	
nG	va à la ligne n
G	fin du fichier

### 4.2.3 Sur la ligne de commande

#### Domaine d'application de la commande

Les commandes peuvent être effectuées sur la ligne courante ou sur une partie du fichier qu'on spécifie au début de la commande.

Commande	Domaine
1,n	de la ligne 1 à la ligne n
%	tout le fichier
n,\$	de la ligne n à la fin du fichier
.	ligne courante

#### Recherche

Il est possible de rechercher dans le fichier une certaine séquence de caractères ou un certain modèle utilisant les expressions régulières (voir Expressions régulières).

Commande	Domaine
/	recherche vers le bas
?	recherche vers le haut

## Substitutions

La commande "s" permet d'effectuer dans un fichier vi des substitutions systématiques, sur un domaine limité comme indiqué précédemment ou sur la ligne courante uniquement. On peut également rechercher des expressions régulières.

Exemple	
<code>%s/57 Metz/57000 Metz</code>	remplace "57 Metz" par "57000 Metz" dans tout le fichier
<code>15s/[iI]nstitut/IAI/g</code>	remplace dans la ligne 15 toutes les occurrences de "institut" ou "Institut" par "IAI".

## 4.3 Expressions régulières

Grâce aux expressions régulières, il est possible d'effectuer des opérations (recherche, substitution) sur des données correspondant à un certain modèle plus complexe que la simple reconnaissance d'une chaîne de caractères.

Caractère spécial	Signification
.	un caractère aléatoire
*	une suite de 0 à n caractères aléatoires
^	négation du caractère suivant
[ - ]	délimitation d'un jeu de caractères
[A-Z]	ensemble des majuscules
/	le caractère suivant doit être pris au sens littéral et non spécial /symbole /* recherche la chaîne de caractères "symbole **"
\( \)	regroupement en une sous-expression
^ (en début de chaîne)	début de ligne
\$ (en fin de chaîne)	fin de ligne

## 4.4 Manipulation de données en utilisant des filtres

La plupart des commandes citées ici s'appliquent par défaut sur l'entrée standard (le texte saisi au clavier) et la sortie standard (l'écran). Si on veut les appliquer à un fichier d'entrée et écrire le résultat dans un fichier de sortie, on peut utiliser le caractère de redirection >.

Commande	Action
<code>grep modèle fichier &gt; sortie</code>	écrit dans <i>sortie</i> toutes les lignes qui répondent au critère <i>modèle</i> qui peut être une expression régulière ("global regular expression print").
<code>grep "03.87" adresses &gt; moselle</code>	écrit dans le fichier <i>moselle</i> toutes les lignes du fichier <i>adresses</i> contenant "03.87". Ne modifie pas le fichier original.
<code>grep -v</code>	commande complémentaire de grep : extrait les lignes qui ne correspondent pas au modèle.
<code>sed 'commande' entrée &gt; sortie</code>	effectue commande sur le fichier <i>entrée</i> , écrit le résultat dans <i>sortie</i> .
<code>sort entrée &gt; sortie</code>	trie les lignes du fichier <i>entrée</i> selon l'ordre ASCII.

<code>sort -u</code>	trie en éliminant les lignes identiques.
<code>cut -d'#' -f 1 entrée</code>	considère le fichier <i>entrée</i> comme un tableau avec # comme délimiteur de colonnes et en extrait le champ 1.
<code>wc fichier</code>	compte les lignes, mots, caractères ("word count") -l compte les lignes -w compte les mots -c compte les caractères.

#### 4.4.1 Fichiers de type base de données

Les fichiers de type base de données sont constitués de colonnes comprenant un type d'information par colonne (par exemple, une liste de noms et d'adresses). Il n'existe pas de réelle fonction tableur dans Unix, chaque colonne est un champ délimité par un caractère spécial (espace, tabulation ou symbole au choix).

<b>Tri</b>	
<code>sort -t"#" -k 1 adresses</code>	trie le tableau <i>adresses</i> , ayant # comme délimiteur de colonne, dans l'ordre croissant du premier champ.
<b>Extraction de champs</b>	
<code>cut -d"#" -f 1 adresses</code>	extrait le premier champ du tableau <i>adresses</i> , ayant # comme délimiteur de colonne.

Il est également possible de fusionner deux fichiers de type base de données. Cette opération nécessite toutefois quelques précautions concernant les deux fichiers. Ils doivent :

- avoir au moins un champ commun,
- être triés selon le même critère,
- avoir le même délimiteur de colonne.

La sortie par défaut est le champ servant de critère de jonction, les champs de fichier1 et ceux de fichier2 pour les lignes appariées.

Prenons ici l'exemple de deux fichiers-tableaux à deux colonnes séparées par le caractère #. Le tableau "*téléphone*" comprend dans la première colonne des noms, dans la seconde des numéros de téléphone. Le tableau "*adresses*" comprend dans la première colonne des noms, dans la seconde des adresses. Pour les deux fichiers, le critère de tri est la colonne des noms.

<b>Fusion</b>	
<code>join -t " #" -1 1 -2 1 téléphone adresses &gt; coordonnées</code>	Écrit dans <i>coordonnées</i> toutes les lignes qui ont pu être appariées, c-à-d. les noms figurant avec un numéro de téléphone dans <i>téléphone</i> et avec une adresse dans <i>adresses</i> . -o permet de spécifier le format de sortie -v affiche en sortie les lignes non appariées.

## 4.5 Enchaîner des commandes

### 4.5.1 Utilisation des "pipes"

Le symbole | représente un "pipe", un tuyau. Il permet d'enchaîner plusieurs commandes sur une même ligne de commande, la sortie de la première commande étant l'entrée de la seconde et ainsi de suite. Les commandes sont effectuées au fur et à mesure que les données sont générées : il n'est pas nécessaire d'attendre que le premier fichier soit généré pour effectuer la deuxième commande, ce qui permet un gain de temps notable lors du traitement de grandes quantités de données.

Exemple
<pre>grep '03.87' téléphone   sort -u &gt; moselle</pre> <p>extrait du fichier <i>téléphone</i> toutes les lignes contenant la chaîne de caractères "03.87", les trie par ordre alphabétique et les écrit dans le fichier <i>moselle</i>. Le fichier original reste inchangé.</p>

### 4.5.2 Scripts

Lorsqu'une suite de commande est souvent effectuée, il est avantageux d'en faire un petit programme, un script. Il est ensuite possible de l'exécuter comme n'importe quelle autre commande, à la condition que le fichier soit exécutable.

#### Permissions d'un fichier

Il est possible d'effectuer sur un fichier trois opérations, symbolisées chacune par une lettre : lecture ("read", r), écriture ("write", w), exécution ("execute", x). Le créateur du fichier définit les permissions aux différents utilisateurs : le créateur lui-même ("user", u), le groupe de l'utilisateur ("group", g) et les autres utilisateurs ("others", o).

Exemple
<pre>chmod u+x,o-w script_</pre> <p>donne à l'utilisateur la permission d'exécution et enlève aux autres le droit d'écriture sur le programme <i>script_</i></p>

#### Exemple de script : do\_grep

<pre># Searches for a regular expression, writes out all matching lines in file .mit, all unmatched lines in file .ohne  #!/bin/csh -f  if (\$#argv != 2) then     echo ""     echo "Illegal number of arguments (\$#argv)"     echo "Usage :\$0 "     echo "        Input file"     echo "        Output file"     echo ""     exit 1 endif  set i = \$1 set o = \$2  echo "" echo "Regular expression to find:"</pre>	<p>Ligne de commentaire</p> <p>Shell utilisé (indication des commandes par défaut)</p> <p>Affichage à l'écran si syntaxe de commande incorrecte</p> <p>Définition des variables i et o</p> <p>Affichage :</p>
---	---

<pre> set r = \$&lt;  grep \$r \$i &gt; \$o.mit grep -v \$r \$i &gt; \$o.ohne  echo " "  echo "=====" echo "Input file:" wc -l \$i echo "Results:" wc -l \$o.* echo "====="  exit </pre>	<p>Attente de saisie d'une valeur pour la variable r</p> <p>Exécution des commandes grep sur le fichier</p> <p>Affichage des résultats :</p> <p>Nombre de lignes du fichier d'entrée Nombre de lignes des fichiers de sortie</p> <p>Fin du programme</p>
--	--

Les scripts permettent d'écrire des programmes sans devoir assimiler un langage de programmation complexe. Ils permettent d'effectuer des opérations répétitives sur de nombreux fichiers sans entrer les commandes manuellement, et donc en réduisant le risque d'erreur dans la syntaxe. Ceci assure une cohérence dans le traitement des données et des résultats fiables.

Unix est, comme le dit Hahn, "facile à utiliser mais difficile à apprendre". L'apprentissage des commandes nécessite un investissement de temps, mais l'utilisation ultérieure du système n'en est que plus rapide. La rigueur requise dans la syntaxe des commandes peut paraître au premier abord un obstacle, mais permet d'effectuer des actions extrêmement précises sur un ou plusieurs fichiers.

Travailler sous Unix permet d'avoir une autre approche de l'informatique que celle des PC. Passer d'un environnement fait d'icônes et de menus à une interface uniquement textuelle donne une petite idée du travail sur les premiers terminaux, à l'époque des débuts de la TA.

## Conclusion

Ce stage m'a permis d'appréhender la Traduction Automatique et le Traitement Automatique des Langues comme des réalités et non plus comme des théories "de rêveurs".

Des applications bien concrètes existent déjà pour les entreprises : MULTIDOC est à ce jour utilisé chez plusieurs constructeurs automobiles (Volvo, Saab, BMW, Renault...) et ce pour différentes langues.

S'il reste une part de rêve dans le TAL, qui sait si le rêve ne sera pas réalité d'ici quelques années ?

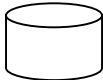
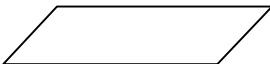


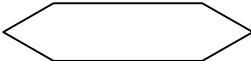
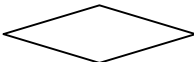

Les autoroutes de l'information sont déjà construites. Certains n'y ont pas accès, bloqués par la barrière de la langue dans ce monde où l'anglais prédomine. Ce problème pourrait n'être plus qu'un vague souvenir lorsque les publications en ligne seront, grâce à un module enconvertisseur et un déconvertisseur UNL, disponible simultanément dans une multitude de langues...

La Traduction Automatique n'est pas non plus au bout de ses ressources. CAT2 est un système de TA qui effectue une analyse morphologique et syntaxique de la langue source et transfère la sémantique de la langue source en langue cible. Il génère ensuite une phrase correspondant aux règles syntaxiques et morphologiques de la langue cible. Ce système de transfert s'approche ainsi au maximum d'un système de pivot tel qu'on en parle en théorie.

Mais d'ici à ce que les machines traduisent de façon acceptable sur un domaine aléatoire, les traducteurs (humains) ont encore de beaux jours devant eux. Ils utilisent déjà des mémoires de traduction pour éliminer l'aspect répétitif de leur tâche et respecter une grande cohérence. Bases de données terminologiques et systèmes notionnels multilingues doivent les aider à atteindre un haut niveau de précision, pour peu que les traducteurs investissent le temps et la réflexion nécessaire à leur construction... Et au moment de baptiser les innovations techniques, on pourra compter sur les outils de gestion de terminologie de l'IAI pour assister le travail d'acceptation terminologique des néologismes.

## Annexes

### A. Légende du diagramme de principe

<b>Types de données</b>	
	Banque de données
	Données
	Document
<b>Procédures</b>	
	Procédé
	Préparation
	Alternative
	Tri

## B. Références

Cet index est thématique, chaque section correspondant à un thème traité dans le rapport.

Un grand nombre de références ne sont pas bibliographiques mais renvoient à une page Internet. Les adresses URL indiquées sont exactes au 20 septembre 2000.

### **Correction de dictionnaire**

*Dictionnaire Universel Francophone*. collaboration de l'AUPELF-UREF et des Editions Hachette.

<http://www.francophonie.hachette-livre.fr/>

Yahoo, moteur de recherche francophone, permet d'effectuer des recherches avancées.

<http://fr.yahoo.com/>

### **Traduction et localisation**

Benito, Daniel et Rico, Celia. 1999. *Déjà Vu - Productivity system for translators - User's manual*. Atril Software.

Schumann, Britta. 2000. *Software-Lokalisierung - Eine Einführung oder: Was unterscheidet den Lokalisierer vom Übersetzer?* Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen. Vol. 16. Fachrichtung 8.6 - Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen - Universität des Saarlandes.

Site de l'IAI.

<http://www.iai.uni-sb.de>

Pages Internet de la section "Traduction Automatique" de l'Université de la Sarre, dirigée par le Professeur J. Haller.

<http://www.iai.uni-sb.de/MT-DEPT/home.html>

Eurodicautom, base de données terminologiques de l'Union Européenne.

<http://eurodic.ip.lu>

Commission des Communautés Européennes. 1997. *Proposition de directive du Conseil relative aux véhicules hors d'usage*. COM (97)358.

<http://www.europa.eu.int/comm/environment/docum/97358-fr.pdf>

Rapport Peugeot Citroën. 1999. *Environnement et automobile*.

[http://www.psa.fr/index\\_enviro.html](http://www.psa.fr/index_enviro.html)

Miquel, Gérard. 1999. *Recyclage et valorisation des déchets ménagers*. Rapport 415 (98-99) - Office parlementaire d'évaluation des choix scientifiques et technologiques. Section E. Les nouveaux créneaux, 1. Les véhicules hors d'usage

<http://www.senat.fr/rap/o98-415/o98-41529.html>

Site donnant des conseils pour l'entretien et la réparation de son véhicule.

[http://www.321auto.com/entretien\\_reparation/glossaire.asp](http://www.321auto.com/entretien_reparation/glossaire.asp)

Site des passionnés de l'automobile, regroupement de clubs d'amateurs.

<http://www.motorlegend.com/wacs/> "Technique"

### **Terminologie**

Arntz, Reiner et Picht, Heribert. 1989. *Einführung in die Terminologiearbeit*. Studien zu Sprache und Technik. Vol. 2. Reiner Arntz und Norbert Wegner. Hildesheim, Zürich, New York: Georg Olms.

Theofilidis, Axel. 2000. Rapports d'analyse de différentes bases de données terminologiques.

Wright, Sue Ellen et Budin, Gerhard. 1997. *Handbook of Terminology Management*. Vol. 1. Basic Aspects of Terminology Management. Amsterdam, Philadelphia: John Benjamins Publishing Company.

ISO 1087:1990. *Terminologie - Vocabulaire*. ISO/TC 37. Première édition.

ISO 12620:1999. *Aides informatiques en terminologie - Catégories de données*. TC 37/SC 3.

ISO 12200:1999. *Aides informatiques en terminologie - Format de transfert de données terminologiques exploitables par la machine (MARTIF) - Transfert négocié*. TC 37/SC 3.

ISO 704:1987. *Principes et méthodes de la terminologie*. TC 37/SC 1.

DIN 2330:1993. *Begriffe und Benennungen; Allgemeine Grundsätze*.

Leclair, Nathalie. Banque de données terminologiques du Service de la langue française du Ministère de la Communauté française de Belgique.

<http://www.cfwb.be/franca/bd/bd.htm>

Base de données de la Délégation générale à la langue française.

Menu "Vocabulaire et terminologie", section "L'enrichissement de la langue française : terminologie et néologie", chapitre "II. Termes, expressions et définitions publiés au Journal officiel", paragraphe "II. La base de données de la DGLF".

<http://www.culture.fr/culture/dglf/>

Van Campenhoudt, Marc. 1997. *Abrégé de terminologie multilingue*. Centre de recherche en terminologie de l'Institut Supérieur de Traducteurs et Interprètes. Bruxelles.

<http://www.termisti.refer.org/theoweb1.htm>

## **Unix**

Abrahams, Paul W. et Larson, Bruce R.. 1996. *Unix for the impatient*. Second Edition. Addison-Wesley Publishing Company.

Hahn, Harley. 1994. *Unix : Guide de l'étudiant*. Ed. Dunod.