VERS L'UTILISATION DES MÉMOIRES DE TRADUCTION POUR LA LOCALISATION DES LOGICIELS : LE PROJET *ALAMET*

Boubaker MEDDEB HAMROUNI[1et2], Laurent FISCHER[1] et Mathieu LAFOURCADE[1]

(1) *GETA - CLIPS - IMAG Campus,* (2)*WinSoft SA., Grenoble, France*

INTRODUCTION

Tous les indicateurs économiques montrent la naissance d'une nouvelle phase de traitement de l'information et de la communication : le multilinguisme. Les échanges commerciaux entre les pays évoluent et les moyens informatiques doivent suivre cette évolution. Les constructeurs et éditeurs des outils informatiques avaient jusqu'à présent vécu, plus ou moins tranquillement, des marchés des pays développés. Aujourd'hui, ce n'est plus vrai, et ces constructeurs doivent viser d'autres marchés potentiels.

Cet article analyse les problèmes théoriques et pratiques — aussi bien terminologiques qu'informatiques — liées à la localisation de logiciels et décrit l'architecture d'un système d'Aide à la Localisation de logiciels pAr MEmoire de Traduction (ALAMET). Cet environnement permet d'automatiser la chaîne de traduction d'un produit d'une langue source vers une nouvelle langue cible. ALAMET constitue aussi une solution à la fois technique et théorique à la normalisation de la terminologie utilisée dans les applications informatiques.

Bien que notre approche n'ait pas été complètement implémentée, nous avons pu l'expérimenter partiellement sur un produit anglais et nous avons réussi à obtenir rapidement un produit localisé en deux langues, sans avoir des connaissances dans la langue cible et surtout en un temps beaucoup plus court qu'avec une approche classique manuelle.

1. LOCALISATION DE LOGICIELS ET MÉMOIRE DE TRADUCTION

La localisation d'un logiciel est l'élaboration d'une version dans une langue spécifique adaptée à un nouveau marché. Cette définition reste très grossière, car la localisation va au-delà de la simple traduction des ressources textuelles contenues dans un

programme. Les aspects culturels peuvent avoir un impact sur certains aspects de l'interface utilisateur. Cependant, nous nous limitons ici aux éléments textuels.

D'une façon générale, la localisation d'un produit nécessite deux étapes distinctes :

- une modification profonde du code de l'application source pour permettre la saisie, l'affichage, l'impression et l'échange des documents créés. Ce qui nécessite principalement des compétences techniques;
- La traduction de l'interface utilisateur (alertes, menus, boîtes...), des manuels, etc. Ce qui nécessite principalement une connaissance de la langue cible et une maîtrise d'un certain nombre d'outils informatiques.

Actuellement, le travail de traduction de l'interface utilisateur se fait 1) d'une manière essentiellement manuelle, 2) sans aucune réutilisation des ressources terminologiques qui pourtant existent dans les applications déjà localisées, et 3) sans aide vis-à-vis des problèmes terminologiques.

La mémoire de traduction est une technique qui vise à accélérer le processus de traduction en faisant intervenir des bases de données contenant les traductions précédemment réalisées. De telles bases s'enrichissent au fur et à mesure des traductions effectuées. Elles sont multilingues et peuvent donc faciliter la traduction d'une langue source vers plusieurs langues cibles simultanément. De plus, elles permettent d'augmenter la cohérence terminologique entre produits et entre traductions.

2. OUTILS EXISTANTS

Dans le cadre de l'internationalisation croissante du marché logiciel, les différentes sociétés éditrices de systèmes d'exploitation (Microsoft, Apple, SUN...) et/ou d'applications bureautiques (Microsoft, Adobe...) sont confrontées à la nécessité d'adapter leurs produits à des marchés extérieurs. C'est à ce prix que de telles sociétés peuvent être présentes sur des marchés en pleine expansion (Asie, Moyen-Orient, Europe de l'Est). Si certaines sociétés ne disposant pas des compétences nécessaires (aussi bien linguistiques que techniques) ont été amenées à sous-traiter ce processus, d'autres ont mis en place des outils leur permettant d'automatiser le processus de localisation de leurs produits.

Plusieurs systèmes d'aide à la localisation ont été développés. En général, ils demeurent spécifiques à leur système d'exploitation d'origine et ne sont pas conçus pour traiter une même application portée sur deux plateformes différentes. On peut citer par exemple :

- Le système I.L.O. (Internationalisation et Localisation de l'Offre) de Bull (1996) qui permet la localisation systématique à partir d'une langue source de tout produit déjà internationalisé dans un certain nombre de langues. Cet outil utilise non pas une simple base terminologique, mais un système de traduction

automatique. Une révision humaine est toujours nécessaire avant la génération du produit localisé.

- Dans le monde Unix, un module (IMT) s'intègre au système d'exploitation et offre des capacités multilingues (saisie, affichage) (Archibald & Darisse 1981). Dans le même domaine, on trouve en outre des utilitaires DVX (Development System Extensions) d'aide à la localisation (extraction des chaînes à partir des applications, outils d'indexation).
- Microsoft propose un ensemble d'outils (RLTools — Resource Localization Tools) qui permettent de localiser des applications Windows (3.11, 95, NT). Ces outils ont été utilisés pour la localisation de toute la gamme des produits Microsoft (Word, Excel, Work...) dans leurs versions multilingues.
- Apple propose une application MacOS (AppleGlot) qui permet de localiser des applications Macintosh d'une manière quasi-automatique et sans l'utilisation d'un éditeur de ressources.

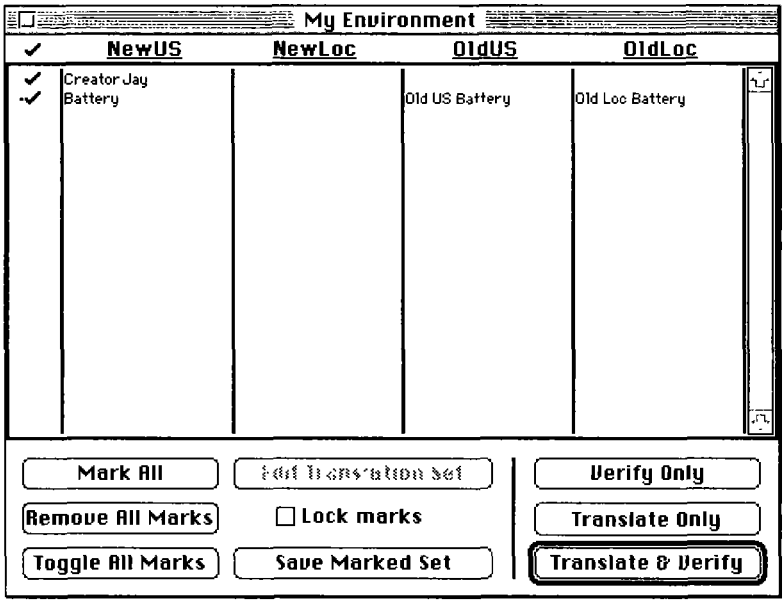
Les outils AppleGlot et RLTools nous ont servi de terrain d'expérimentation.

2.1 AppleGlot (Apple Computer Inc.)

AppleGlot permet de localiser un produit ou un ensemble d'applications Macintosh. Cette application permet non seulement de localiser une application mais donne la possibilité de mettre à jour une application déjà localisée, s'il y a eu de nouvelles versions, en se basant sur un fichier d'historique.

La procédure de localisation commence par la création d'un dossier dans lequel l'application dans la langue source doit être placée. Après le démarrage d'AppleGlot, l'écran suivant apparaît :

- NewUS : contient l'application en langue source (par défaut on suppose qu'elle est en anglais).
- NewLoc : contiendra l'application localisée dans la langue cible.
- OldUS : sert pour une mise à jour et contient l'ancienne version de l'application en langue source.
- OldLoc : contient l'application dans sa précédente version déjà localisée dans la langue cible.



Le bouton «Translate Only», permet d'extraire toutes les chaînes de l'application source. Ces ressources textuelles sont représentées dans le format suivant :

```
{Le texte ici entre accolades est un commentaire}
DITL 128 dialog item text 8 (1) {N'a pas trouvé de traduction}
<File>
<> {entrer la traduction «Fichier» ici }

DITL 128 dialog item text 26 (3) [?? Text Match - Internal Guess]
<Print>
<Imprimer>
{ici il a trouvé en se basant sur une traduction déjà faite dans une version antérieure}
DITL 128 dialog item text 62 (7)
<Settings>
<>

DITL 129 dialog item text 26 (3) [??:•• Wild Guess by position]
<Paper>
< >

STR 128 text 1
<Some Text>
<>
```

Il est nécessaire de saisir manuellement les traductions dans ce fichier texte au niveau de la seconde ligne (entre < et >) et on rappelle le dialogue. «Translate & Verify» permet alors d'injecter les traductions et de générer l'application cible.

L'atout d'AppleGlot est sa simplicité d'utilisation, sa capacité à extraire toutes les chaînes, même si elles ne font pas partie des ressources standards du Macintosh (Str#, MENU...). En plus, la possibilité de mettre à jour une application déjà localisée constitue une fonctionnalité importante. En fait, cet outil gère d'une manière indirecte et implicite une mémoire de traduction.

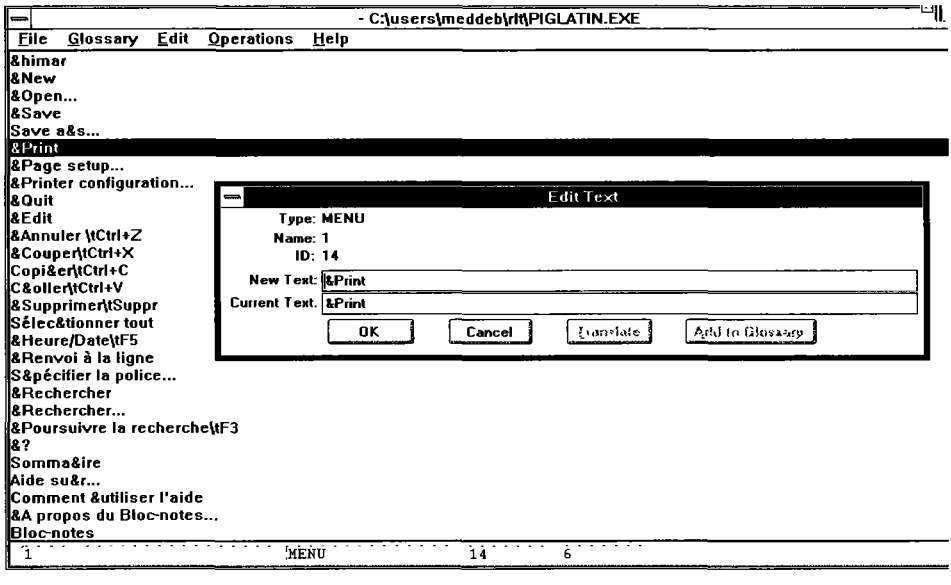
Par contre, tout comme les outils Microsoft, il n'est pas possible de faire les traductions d'une manière automatique, c'est-à-dire qu'on ne peut pas intégrer une mémoire de traduction à l'application pour faire une traduction automatique quand il s'agit de localiser une nouvelle application.

2.2 RL Tools (Microsoft Corp.)

Les outils proposés par Microsoft peuvent localiser un produit, un produit et ses accessoires (filtres, aide...) ou un ensemble de produits. L'outil central (RLQuikEd) existe sous forme d'une application GUI (Graphic User Interface) ou d'un ensemble de commandes DOS. RLQuikEd permet de charger une application et d'en extraire toutes les chaînes. Il propose alors à l'utilisateur une table dans laquelle il doit rentrer la traduction d'une chaîne en regard de la chaîne proposée dans la langue source. Une fois toutes les chaînes traduites, une opération de génération de l'application cible peut être lancée afin d'obtenir le produit localisé.

D'autres outils permettant de réajuster la taille des dialogues, de créer des projets de localisation (nécessaire lors de localisation de plusieurs applications), sont également proposés dans le kit RLTools.

RLTools n'est en fait, qu'un outil d'extraction des chaînes et d'injection du résultat de la traduction. Aucune mémoire de traduction n'est disponible. Il est possible de créer et de réutiliser une mémoire de traduction dans un format texte simple (entrée en langue source, tabulation, entrée en langue cible). Néanmoins, il n'existe pas d'outils de gestion de la mémoire de traduction et aucune possibilité n'est offerte pour suggérer des mots voisins. Par exemple, si on cherche à traduire la chaîne «File:» et si dans la base on a le couple «(File, Fichier)», le système est incapable de fournir la traduction à cause du caractère «:».



2.3 Analyse

Au vu des deux outils présentés ci-dessus, on peut faire un certain nombre de remarques :

- Ces deux outils constituent de bons extracteurs des chaînes à traduire et de bons injecteurs des chaînes déjà traduites.
- Ils n'utilisent pas de mémoires de traduction externes, ou pas efficacement.
- Ils ne permettent pas de faire d'autres actions de localisation (symétrisation des dialogues, réajustement des boutons...).
- Quand ils utilisent des mémoires de traduction, la recherche des chaînes n'est pas très puissante. On aimerait traduire «O&pen» par «O&uvrir» si le dictionnaire contient «Open» et «Ouvrir».
- Ils sont mono-plateforme. Une application Macintosh, bien qu'identique au niveau de son interface sur Windows, se localise avec AppleGlot sur Macintosh et RLTools sur Windows.

Ce dernier point est particulièrement critique. En effet, il est souhaitable de disposer d'un environnement de localisation dont le fonctionnement n'est pas lié à un système d'exploitation ou à un constructeur particulier. D'une façon générale, seule la phase d'extraction et de réinsertion des ressources textuelles reste fortement dépendante de l'origine des produits à localiser.

3. LE PROJET ALAMET

Le projet vise à définir d'une part une architecture générale d'un système de localisation générique, et d'autre part un format de fichier (nommé TRADE) permettant de représenter les unités textuelles en entrée et en sortie de la mémoire de traduction.

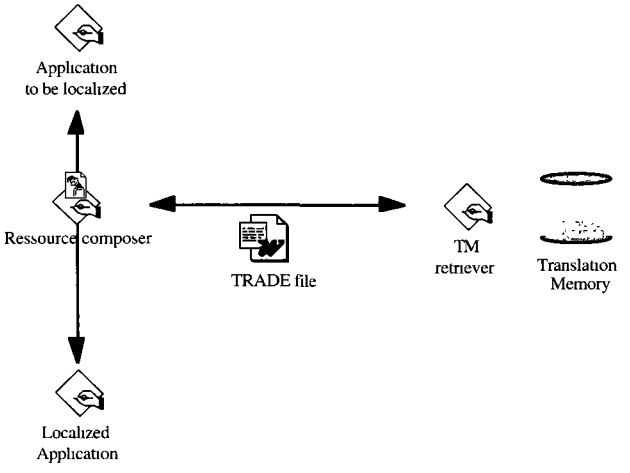
3.1 Architecture

L'architecture du système ALAMET est centrée autour du format de fichier TRADE (dont une description est donnée ci-après). Nous considérons deux principaux composants dans l'architecture ALAMET : 1) le composeur de ressources et 2) le gestionnaire de mémoire de traduction.

Si on dispose déjà d'une mémoire de traduction, le processus de localisation se décompose de la façon suivante :

- extraire les ressources textuelles de leurs applications sous la forme d'un (ou plusieurs) fichier(s) TRADE (nommé TRADE.in);
- «nourrir» le gestionnaire de mémoire de traduction avec ce fichier;
- obtenir un fichier TRADE (TRADE.out) contenant les ressources lexicales avec leur(s) équivalent(s);
- réviser le fichier TRADE.out. Ceci implique d'une part de vérifier les équivalents proposés, d'en sélectionner un seul pour chaque ressource textuelle, et de compléter ceux qui n'auraient pas été trouvés. Nommons ce fichier TRADE.rev;
- injecter le fichier TRADE.rev dans le composeur de ressources afin d'obtenir les applications localisées dans les langues cibles;
- injecter le fichier TRADE.rev dans le gestionnaire de mémoire de traduction afin d'augmenter la mémoire de traduction (cette étape est optionnelle, mais néanmoins cruciale pour l'amélioration de l'efficacité du système).

Si le système est vierge de toute mémoire, l'étape consistant à trouver les équivalents doit être faite à la main. Toutefois, en pratique, on cherchera à alimenter dès le départ la mémoire de traduction avec des couples d'applications déjà localisées.



Le composeur de ressources

Le composeur de ressources permet l'extraction et la réinsertion des ressources textuelles. Précisément, ce composant doit être capable :

- d'extraire les ressources textuelles d'une application et d'en produire un fichier TRADE;
- de produire une application localisée, à partir d'un fichier TRADE disposant des traductions complétées.

Le processus d'extraction des ressources textuelles dépendant fortement du système d'exploitation et/ou du développeur de l'application, le composeur prévoit un mécanisme d'extension («plug-ins»). Ces extensions permettent de décrire ce processus d'extraction selon les applications.

Le gestionnaire de mémoire de traduction (GMT)

La tâche majeure du gestionnaire de mémoire de traduction est de compléter un fichier TRADE.in avec les équivalents trouvés dans la mémoire de traduction.

Dans le premier prototype d'ALAMET (en cours de développement), la mémoire de traduction ne propose qu'un mode de correspondance exacte pour la recherche des chaînes. Nous cherchons à intégrer des modes de recherche plus «flous» incluant :

- la gestion de mots-clés (origine de l'application, version, etc.) avec des niveaux de priorité;
- la recherche de chaînes en correspondances partielles.

Le GMT accepte de recevoir en entrée un fichier TRADE en vue de l'augmentation de la mémoire de traduction. De plus, nous envisageons d'accepter des fichiers de dictionnaires terminologiques comme ceux développés dans le cadre des projets FE*.

3.2 Le format de fichier TRADE

Le format TRADE permet de décrire un ensemble de ressources textuelles (en langue source) associées à leur(s) équivalent(s) dans plusieurs langues cibles. Ce format inclut les informations suivantes :

- description globale de l'origine des ressources textuelles (système d'exploitation, applications, version, etc.);
- informations permettant la réinsertion des ressources textuelles en langage cible dans les applications en versions localisées;
- spécification des codages ou des transcriptions utilisés pour les contenus textuels.

De plus, ce format est conçu de façon à être suffisamment lisible par les personnes en charge de la localisation.

Informations d'extractions

Tous les outils existant actuellement dépendent directement des systèmes d'exploitation pour lesquels ils ont été conçus. Nous souhaitons adopter une démarche générique qui rende le composeur de ressources complètement indépendant des systèmes d'exploitation et des éditeurs.

Les informations d'extraction permettent au composeur de ressources de réinsérer les ressources textuelles dans les applications (afin de produire les versions localisées). Il s'agit d'une simple chaîne de caractères. Les informations contenues dépendent directement de l'application à localiser.

Par exemple, sous MacOS on pourra avoir :

```
<Rsc-descriptor>
DITL 128 dialog item text 62
</Rsc-descriptor>
```

Les extensions du composeur permettent de générer ces informations en fonction du type d'application concernée. Elles gèrent également le processus inverse de réinsertion de ressources en fonction de ces informations.

Informations textuelles

Les informations textuelles concernent plus directement la chaîne de caractères à localiser. En plus de la chaîne elle-même, nous considérons les informations de formatage et de codage/transcription.

Le formatage gère l'apparence du texte (gras, italique, police, etc.). Sous Windows, le raccourci clavier est indiqué à l'aide du soulignement d'une lettre particulière. Par exemple dans le menu d'une application Windows, on trouve l'article «File» qui permet d'activer la commande d'ouverture d'un fichier en appuyant sur la touche Alt+F. Il est nécessaire de garder ce type d'information lors de la localisation de l'application.

Le code langue indique la langue ou la famille des langues utilisées. En pratique, cette information est rarement explicitée dans les ressources textuelles des applications. Cependant, il est possible d'inclure cette information comme un des paramètres d'extraction.

Le codage indique le type de codage de la chaîne de caractères (par exemple MacRoman, Unicode, ASCII, ASMO449, etc.).

La transcription indique le type de transcription utilisée pour la représentation des chaînes. Cette information est optionnelle.

Informations générales

Les informations générales sont représentées sous forme de variables à valeur ensembliste (énumérées) indiquant des informations comme le système d'exploitation, l'application, etc. Certaines de ces informations peuvent être indiquées globalement. Toutes peuvent être locales à chaque élément textuel.

Nous avons par exemple :

- OS : MacOS, Win95, Win311, ...
- Application : Pagemaker, MSWord, ...
- Version : 1.0, 2.1, ...
- Type : alerte, dialogue, menu

D'autres types de variables peuvent être envisagés.

Le gestionnaire de mémoire de traduction essaye de trouver l'équivalent qui satisfait au mieux les contraintes qu'imposent ces mots clés.

Exemple

```
<TRADE>

<HEAD>
<Source-language>English<\Source-language>
<Coding>MacRoman<\Coding>
<Application> Eudora <\Application>

<\HEAD>

<TRADE-item>

        <Rsc-descriptor>DITL 128 dialog item text 62<\Rsc-descriptor>
        <Source>
        <String>Don't trash<\String>
        <Application>Eudora<\Application>
        <OS>MacOS<\OS>
        <Version>2<\Version>
        </Source>

        <Targets>
                <Target-language>French<\Target-language>
                <String>Ne pas jeter<\String>
                <String>Ne pas détruire<\String>

                <Target-language>Arabic<\Target-language>
                ...
        <\Targets>

<\TRADE-item>

<TRADE-item>
        <Rsc-descriptor>DITL 128 dialog item text 63<\Rsc-descriptor>

        <Source>
        <string>Cancel<\string>


        <\Source>
        <Targets>
                <Application>Eudora<\Application>
                ...
        <\Targets>
<\TRADE-item>

<\TRADE>
```

Comme dans le format HTML, les retours chariot ne sont pas significatifs. Les blancs, tabulations, etc. entre les étiquettes ne sont pas significatifs.

4. ALAMET — UNE PREMIÈRE EXPÉRIENCE

Dans le cadre des projets de dictionnaires Français-Anglais-Malais et Français-Anglais-Thaï soutenus par le MAE, le GETA a acquis une expérience dans la réalisation et la mise à disposition de dictionnaires multilingues (une langue cible vers plusieurs langues sources). Ces projets ont permis d'arriver à une méthodologie efficace de constitution de tels dictionnaires en utilisant des logiciels du commerce.

La société WinSoft a depuis longtemps une expérience dans les logiciels multilingues et dans le service de localisation. Ses clients sont, entre autres, de grandes sociétés de logiciels comme Adobe, Microsoft ou Apple. L'apport de WinSoft consiste essentiellement en deux points :

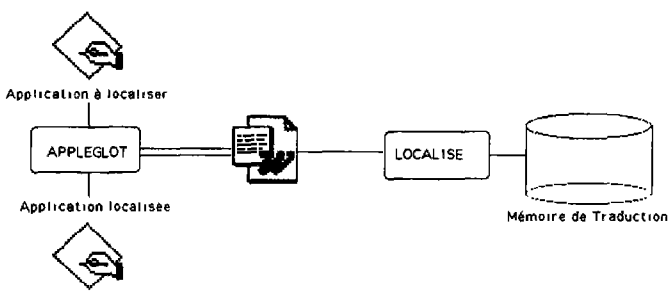
- modifier les logiciels d'origine, pour leur permettre de fonctionner dans des langues utilisant des systèmes d'écriture non romains;
- traduire les ressources de ces logiciels afin de produire des applications localisées pour des marchés comme ceux du Moyen-Orient ou de l'Europe de l'Est.

WinSoft reçoit une demande croissante pour la localisation de logiciels dans des délais courts. Pour ce faire, nous avons conçu une première chaîne de localisation, inspirée de l'architecture d'ALAMET. Nous avons implémenté sous MacOS une première maquette et l'avons expérimenté sur un nouveau produit (PhotoShop 4.0).

4.1 Description

Nous avons utilisé comme outil d'extraction et d'injection le logiciel AppleGlot 2.4 (Apple Computer) et comme outil de gestion de la mémoire de traduction l'application 4D 5.0.2 (ACI).

Notre chaîne de localisation se présente comme suit :



Pour localiser une application, on procède en trois étapes :

1) Utiliser AppleGlot sur l'application source. C'est l'étape d'extraction des chaînes. Cette étape produit un fichier «TEXT» contenant toutes les chaînes à traduire. Ce fichier est dans le format suivant :

```
DITL 128 dialog item text 62 (7)
<Paramètres>
<>

DITL 129 dialog item text 26 (3) [??:•• Wild Guess by position]
<Papier>
< >
```

Les chaînes à traduire sont placées entre les signes «<» et «>» et les chaînes traduites devront être placées dans le «< >» qui suit.

2) Utiliser l'application LOCALISE que nous avons développée. Elle prend en entrée un fichier produit par AppleGlot et cherche les traductions dans la base 4D. Le résultat est un fichier «APPLEGLOT» dans lequel ont été placées les traductions (si elles existent dans la base).

```
DITL 128 dialog item text 62 (7)
<Setting>
<Paramètres >

DITL 129 dialog item text 26 (3) [??:•• Wild Guess by position]
<Paper>
<Papier>
```

L'application LOCALISE est présentée ci-après.

3) Ré-exécuter «APPLEGLOT» sur le fichier produit à l'étape 2 et sur l'application source. C'est l'étape d'injection des ressources. Cela a pour effet de remplacer les chaînes de l'application en langage source par leurs équivalents dans la base, au regard du fichier des traductions et de générer l'application cible dans la nouvelle langue.

4.2 L'outil LOCALISE

LOCALISE a été développé en vue de deux objectifs :

- gérer une base de données terminologiques contenant les termes informatiques les plus utilisés dans les applications Macintosh.

- utiliser la mémoire de traduction pour localiser de nouvelles applications ou mettre à jour des applications déjà localisées.

La base que nous avons construite contient les entrées en 6 langues (français, anglais, arabe, hébreu, japonais, russe) et contient les termes informatiques les plus utilisés dans les interfaces des applications Macintosh.

Cet outil a été écrit avec le langage de programmation de 4D (v5.0.2). Les différentes fonctionnalités relèvent de l'utilisation ou de l'administration :

Utilisation :

Localise permet de lancer la localisation d'une application à travers son fichier de chaînes (glossaire) préalablement produit par AppleGlot. Au lancement de l'action «Localise», l'écran suivant apparaît. L'utilisateur y spécifie :

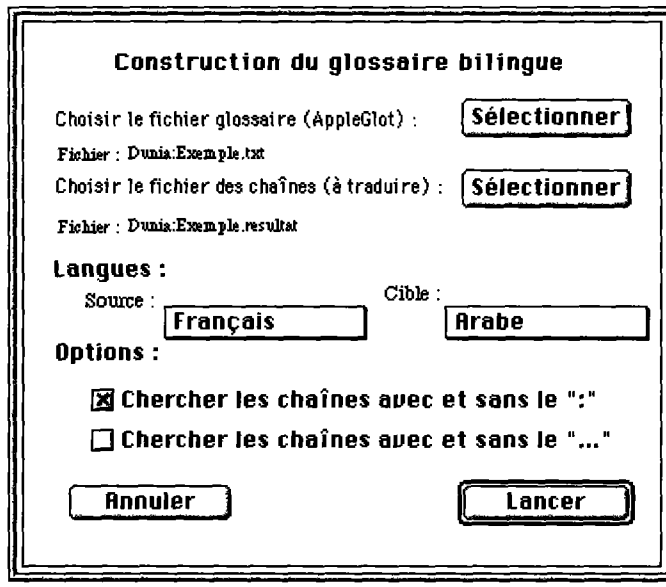
1) le fichier glossaire qui contient les termes dans la langue source,
2) le fichier dans lequel se trouveront les chaînes résiduelles qui ne seront pas traduites (car elles sont inexistantes dans la base) et qui seront traduites par un traducteur humain,
3) la langue source et la langue cible, puisque la base contient des termes multilingues.

Deux options sont enfin proposées :

a) chercher des chaînes avec ou sans le caractère «:»;
b) chercher des chaînes avec ou sans le caractère «…»[1].

Ces deux dernières opérations permettent de trouver le maximum de chaînes.

[1] Le «…» est un caractère ayant son propre codage sur Macintosh.



À la fin de l'opération, un nouveau fichier (ayant le même nom mais avec une extension) est produit et contient les termes traduits.

Consulter la base permet d'afficher le contenu de la base selon un ordre alphabétique spécifié via un pop-up menu.

Rechercher permet de rechercher des termes dans la base.

Administration :

Les fonctionnalités proposées permettent de supprimer, ajouter, modifier, importer, exporter des termes. Ce sont les opérations classiques qu'on trouve dans la plupart des gestionnaires de bases de données.

4.3 Résultats

Cette chaîne de traduction a été utilisée pour localiser vers l'arabe la nouvelle version de PhotoShop à partir d'une version préalablement localisée, PhotoShop 3.0.5. Grâce à notre méthodologie, nous avons réussi d'une part à récupérer toute la terminologie de la précédente version d'une manière automatique, et d'autre part à trouver environ 30 % des nouvelles chaînes dans notre mémoire de traduction. C'est un résultat intéressant, puisque les nouveaux termes sont ceux inhérents au domaine du traitement d'images, ce qui donne a priori peu de chances de les trouver dans notre base.

Il est important aussi de noter que l'opération de localisation a pris beaucoup moins de temps qu'une approche manuelle (déjà réalisée par la société qui édite le logiciel, Adobe) puisque nous avons réussi à récupérer automatiquement les chaînes existantes dans la précédente version et 30 % des nouvelles chaînes lors du temps d'exécution du programme d'injection, soit environ une heure.

Enfin, après cette première expérience, notre mémoire de traduction s'est enrichie de nouveaux termes issus de l'application PhotoShop 3.0.5.

CONCLUSION

Les objectifs du projet ALAMET sont très ambitieux. Mettre en place une chaîne de traduction multi-plateformes est un projet de grande envergure. Aucun système de ce type n'est disponible, ce qui fait que ce projet s'aventure en terrain relativement vierge.

Dans cet article, nous avons seulement posé le problème, décrit les outils existants, et proposé une première solution. Bien que spécifique, elle nous a permis de nous rendre compte de tous les problèmes qu'on peut rencontrer, et a montré concrètement les avantages de notre méthode.

Le format TRADE et l'architecture du système proposé permettent, pour la première fois semble-t-il, de prendre en compte tous les problèmes posés par la localisation automatique.

Pour la suite, nous voulons intégrer dans la structure de la mémoire de traduction des concepts intéressants, comme les bases lexicales par acceptions, définies dans Sérasset (1994) et implémentées par Blanc (1995) dans PARAX.

RÉFÉRENCES

Apple Computer Inc. (1996) : *«AppleGlot», User manual*, CD Tool Chest, Developper CD series, août 1996.

ARCHIBALD, J. & A. DARISSE (1981) : *A Guide to Multilingual Publishing*, DC : Society for Technical Communication.

BLANC, É. (1995) : «Une maquette de base lexicale multilingue à pivot lexical : PARAX», *Lexicomatique et dictionnairiques*, Actes des IV^{es} Journée scientifiques du réseau LTT, Lyon, France, septembre 1995, Beyrouth, FMA et AUPELF-UREF, pp. 43-58.

BULL, S.A. (1990) : «Lignes directrices pour le développement des produits internationaux», *Manuel de présentation du système ILO*, Bull S.A., CEDOC, février 1990.

Microsoft Corp. (1996) : *«RL Tools», Users Manual*, ftp.microsoft.com/pub/Dev/.

SEMMAR, N. (1995) : *Méthodologie de localisation des applications informatiques multimédia*, Nouvelle thèse, Université Paris VI, octobre 1995, 196 p.

SÉRASSET, G. (1994) : *SUBLIM : un Système Universel de Base LexIcales Multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*, Nouvelle thèse, UJF, décembre 1994, 194 p.

PLANAS, E. & Ch. BOITET (1997) : «Towards an Evolution of Memory Based Translation Systems», *PACLING'97*, Meisei Universioty, Ohme, Tokyo, September 1997, 11 p.

QUAND LES MOTS PERDENT LA MÉMOIRE

Muriel CORET

Université Paris 7 - Denis Diderot, Paris, France

1. LA MÉMOIRE MORPHOLOGIQUE

On peut parler d'une mémoire des mots. Ou plutôt de mémoire de la langue qui conserve certaines informations concernant les mots. Je m'intéresserai ici à la mémoire qu'on pourrait appeler *morphologique.*

Comment rendre compte de l'abandon de *extradation,* recensé dans le *Vocabulaire des mots nouveaux* de Mercier (1801) au profit de *extradition*[1], forme attestée dans les dictionnaires du XIX^{e} siècle ? Si la première forme proposée peut sembler parfaitement régulière dans un premier temps (comparons *extrader, extradation* à *réparer, réparation...*) la seconde, elle, a l'avantage de garder en mémoire l'origine du mot par la présence du *-i-* de la série *trado, tradere, traditum.* Le choix de la forme *extradition* a donc l'avantage de rappeler l'origine du mot : ce choix est la manisfestation de l'activité d'une mémoire.

Ce phénomène n'est pas un cas isolé. De nombreux aspects du fonctionnement du lexique révèlent que la langue conserve en mémoire certaines propriétés morphologiques du système latin, qu'elle peut réutiliser en français moderne pour de véritables créations. Il en est ainsi, par exemple, du phénomène de l'allongement thématique, observable dans le lexique du français et défini par Huot (1994) comme l'allongement d'une racine lexicale en *-(a/i)-t/s-* formant une base apte à certaines dérivations. C'est bien aux modèles d'alternances latines *amare / ama-t-um, mittere / miss-um, capere / cap-t-um...* que renvoient par exemple les alternances qu'on observe dans *aim-er / amat-eur, permett-re / permiss-ion, recev-oir / récept-ion...*

Mais comme toute mémoire, celle de la langue connaît des défaillances. Ce sont ces pertes de mémoire que je vous propose d'illustrer ici avec le cas des mots en *-ion.*

2. MÉMOIRE DES MOTS ET FORMATION DES NOMS EN *-ION*

La formation des mots en *-ion* constitue un bon exemple de cet espace de mémoire, en ce sens qu'elle est fondamentalement calquée sur la morphologie latine. On

1 L'exemple est emprunté à M. Glatigny, «Néologisme et lexicographie», Forum de Morphologie, Lille, avril 1997.

observe une grande régularité en ce qui concerne la morphologie de ces noms : le suffixe *-ion* sélectionne des bases dites «thématiques», comportant dans la grande majorité des cas un allongement du radical en *-(a/i)t-*. Il en est ainsi dans *programm-at-ion, répét-it-ion, attribu-t-ion...*, qui comportent un allongement, mais aussi dans *production, concession, adhésion...* qui ne présentent pas en surface un véritable allongement mais dont les bases doivent être considérées néanmoins comme des bases thématiques[2], s'opposant aux formes non thématiques attestées dans les verbes correspondants *produi-re, concéd-er, adhér-er...*

On retrouve là un trait de la morphologie du latin : la base qui apparaît dans le mot en *-ion* est directement calquée sur la formation latine des noms en *-io, ionis*, dans lesquels on retrouve le thème du supin. Exemples :

fr. *N-ion*	**lat. *N-io, ionis***	**Supin**
cession	*cessio*	*cessum*
gestion	*gestio*	*gestum*
correction	*correctio*	*correctum*

Celle qui apparaît dans le verbe est issue, par dérivation savante ou populaire, du radical de l'infectum correspondant :

fr. *V-er*	**V latin**
céder	*cedere*
gérer	*gerere*
corriger	*corrigere*

97 % des mots en *-ion* du Robert électronique[3] relèvent de ce fonctionnement : une même racine se présente sous une double forme, opposant une base thématique pour la suffixation en *-ion* (mais aussi en *-if, -oire, -ure...*) et une base non thématique pour le verbe correspondant (et la suffixation en *-age, -ment...*).

Or, on est amené à isoler un ensemble de mots qui résistent à cette analyse : ils comportent des bases qui, contrairement au cas général, ne présentent aucune variation. Ainsi, à la série régulière *programm-at-ion / programm-er, répét-it-ion / répét-er, attribu-t-ion / attribu-er* s'opposent des mots en *-ion* présentant le même radical que les verbes qui leur sont apparentés, tels que *abras-er / abras-ion, incis-er / incis-ion, invent-er / invent-ion...* au lieu de l'opposition «base thématique + *ion*» *vs* «base non thématique + *er*».

88 couples de mots, *N-ion* et *V-er*, sont dans ce cas. Ils sont regroupés, par famille, dans les 44 entrées du tableau donné ci-dessous en annexe. La totalité des couples *«V-er / N-ion»* de même base y figurent, exhaustivement, accompagnés d'un certain nombre d'informations commentées dans la suite du travail.

Ces mots posent problème. Que s'est-il passé ? La langue dispose-t-elle de deux manières de construire les noms en *-ion* ? Faut-il considérer que ces mots en *-ion* relèvent d'une autre règle que celle décrite plus haut (et donc en restreindre la généralité) ?

2 Ce point ne sera pas développé ici. Cf. Coret (1994).

3 Le *Grand Robert électronique* comporte 2951 mots en *-ion*.

3. RECONSTRUCTION DE VERBES ET PERTE DE MÉMOIRE

Mon analyse s'appuie sur des arguments d'ordre historique, sémantique et morphologique. Elle a permis de mettre à jour les caractéristiques suivantes :

Aspect historique

Une vérification rapide montre que les noms en *-ion* sont apparus relativement récemment dans les dictionnaires (et donc, nous l'admettrons, dans la langue) mais sont généralement donnés comme antérieurs aux verbes correspondants. Un tiers environ des mots considérés ne figurent pas du tout dans le dictionnaire Richelet. On peut mentionner à titre d'exemple :

	Datation[4] du *N-ion*	**Datation du *V-er***
collect	XIV^{e}	XVI^{e}
concoct	1528	XX^{e}
digress	XII^{e}	1838
édit	XIII^{e}	1878
excrét	XVI^{e}	1836
inject	XIII^{e}	XVIII^{e}
inspect	1290	1781
lés	XII^{e}	1611
réfract	XVI^{e}	1752
révuls	XVI^{e}	1845
transit	XIII^{e}	1832
vis	XII^{e}	1668

Dans ce domaine bien sûr, une uniformité totale n'est pas possible. Mais, comme le souligne déjà Huot (1997), l'important est que, majoritairement, ces *N-ion* ne peuvent pas être considérés comme issus des verbes en *-er* et que, d'autre part, la formation des *V-er* est encore un processus vivant aujourd'hui (comme l'attestent quelques formations récemment apparues dans les dictionnaires comme *compacter, collecter...*).

Aspect sémantique

On constate que le sens des verbes formés sur la base des noms en *-ion* est décrit de manière très générale par les dictionnaires à l'aide de la paraphrase[5] «faire / commettre / procéder à [l'action exprimée par le *N-ion*]» — ce qui n'est pas le cas pour les autres couples *V-er* et *N-ion*, pour lesquels au contraire, le verbe sert à définir le nom (paraphrases en «action de V»). On relève par exemple[6] dans le *Petit Robert* :

4 Les datations sont celles de Littré (1863) ou du dictionnaire *Le Petit Robert*.

5 Les dictionnaires montrent sur ce point une telle hétérogénéité que la moindre régularité mérite toute notre attention !

6 Il est à noter que lorsque la définition ne mentionne pas le *N-ion*, elle a généralement recours à une autre unité comportant la base thématique. Cf. *collecter* : «faire une collecte»; *compacter* : «rendre compact»; *exempter* : «rendre exempt»; *professer* : «enseigner en qualité de professeur»; *propulsion* : «faire avancer à l'aide d'un propulseur»...

abraser : «User (une matière, un objet) par **abrasion**»
agresser : «Commettre une **agression**»
confesser : «Entendre (un fidèle) en **confession**»
connecter : «Unir par une **connexion**»
décompresser : «Cesser ou diminuer la **compression** de»
déconnecter : «Supprimer la **connexion** de»
décontracter : «Faire cesser la **contraction** musculaire de»
diffracter : «Produire la **diffraction** de»
digresser : «Faire des **digressions**»
exciser : «Enlever par **excision**»
excréter : «Évacuer par **excrétion**»
exploser : «Faire **explosion**»
imploser : «Faire **implosion**»
impulser : «Donner une **impulsion** à»
objecter : «Opposer (une **objection**) à»
opiner : «Dire, énoncer son **opinion**»
perfuser : «Pratiquer une **perfusion** sur»
préciser : «Apporter des **précisions**»
presser : «Exercer une **pression**»
rééditer : «Donner une nouvelle **édition** de»
réfracter : «Faire dévier par le phénomène de la **réfraction**»
régresser : «Subir une **régression**»
révulser : «Faire affluer par **révulsion** le sang de»
sécréter : «Produire (une substance) par **sécrétion**»
transfuser : «Soumettre à une **transfusion** sanguine»

Ces définitions mettent en évidence une valeur particulière pour ces verbes qu'on pourrait caractériser de verbes «d'effectuation». Ce fait, déjà souligné dans Coret (1994) et Huot (1994), mériterait d'être étudié dans le cadre d'une hypothèse plus générale sur la valeur du segment d'allongement lui-même, qui *n'exprimerait rien d'autre*, selon Huot (1994 : 55) *que l'accompli.*

Aspect morphologique

Ces noms, s'ils peuvent paraître dans un premier temps mal formés en français, sont parfaitement réguliers en regard du modèle latin. Ils présentent une base thématique, issue du thème du supin latin qui, dans presque tous les cas[7], alternait avec une forme simple, conformément aux règles morphologiques du système. Cette alternance figure dans la troisième colonne du tableau en annexe. Citons par exemple :

Français	**Supin latin**	**Infinitif latin**
ablat-	*ablatum*	*auferre*
abras-	*abrasum*	*abradere*
accept-	*acceptum*	*accipere*
agress-	*aggressum*	*aggredi*

7 Ce n'est pas le cas en effet pour *annexio, communio, complexio, opinio, rebellio, recensio,* qui n'ont qu'une base. Ce sont là les seuls cas où on ne puisse pas reconsttruire une base non thématique.

On peut voir qu'il s'agit de formations régulières à l'origine, et on retrouve bien ici l'application de la règle générale citée plus haut : le *N-ion* est formé sur une base thématique, qui alterne avec une base simple. La particularité que je veux pointer ici, c'est que ces bases n'ont subsisté que sous leur forme thématique.

Pour autant, la base non thématique (de formation populaire ou savante) n'est pas totalement éliminée : dans 38 cas sur 44, une base non thématique était accessible pour la formation du verbe et rien, du point de vue morphologique, n'empêchait sa réactivation. Le lexique présente en fait différentes traces de son existence :

- il arrive qu'on retrouve un verbe formé sur la base non thématique dans un état de langue ancien. Ainsi *affaiter, colliger, expeller, invertir, objicer, transir* (face à *affect-, collect-, expuls-, invers-, object-, transit-*) ont été attestés en français et sont mentionnés dans des dictionnaires anciens ou des dictionnaires historiques.

- la base non thématique a pu aussi subsister de manière isolée sous la forme d'un nom ou d'un adjectif encore attesté en français moderne : *diffract- / **diffring**-ent, divis- / **divid**-ende, incis- / **incid**-ence...*

- la base non thématique a pu enfin se maintenir en français moderne, sous la forme d'un premier verbe (normalement constitué), qui se trouve donc en concurrence avec le *V-er* sur base thématique récemment formé. C'est ainsi qu'on a les concurrents

  cueillir (de *colliger*) / *collecter*
  comprimer / compresser
  dire / dicter
  repousser / répulser
  voir / viser...

  L'existence d'un premier verbe (construit sur base non thématique) n'a pas bloqué la création d'un second verbe. Ils sont issus de la même racine et on pourrait considérer que, à un certain niveau d'analyse, ils ont le «même» sens. Pourtant, leur coexistence même empêche de les considérer comme des synonymes. En fait, une fois le premier verbe entré dans le lexique, il acquiert un sémantisme particulier, qui l'éloigne de la simple relation au *N-ion*. C'est cette rupture que pallie la création d'un nouveau verbe, sur la base exacte du *N-ion*.

Dans les cas où on ne peut pas retrouver cette base non thématique dans un mot attesté, il est possible de reconstruire une base possible (que nous noterons par le symbole °) d'après le modèle général d'évolution de la langue, en prenant éventuellement appui sur d'autres séries de mots, qui ont, elles, conservé les deux termes de l'alternance. On pouvait ainsi attendre °*concuire* (d'après *cuire*), °*concroître* (*croître*), °*contraire* ou °*contirer* (*traction / traire, tirer*), °*détéger* (*protection / protéger*), °*diffondre* (*fusion / fondre*), °*excider* (*décision / décision*), °*invenir (prévention / prévenir*)...

Seuls se présentent en fait comme cas particuliers les bases *annex-, commun-, complex-, opin-, rebell-, recens-,* pour lesquelles on ne peut pas reconstruire une base non thématique : elle n'est attestée dans aucun mot du français et ne peut pas être reconstruite d'après le supin puisque en latin déjà, la même base était attestée pour la formation du supin et de l'infinitif correspondant.

4. MÉMOIRE MORPHOLOGIQUE ET RÉGULARITÉ LEXICALE

Ces observations montrent qu'on peut rendre compte de cet ensemble de cas sans remettre en cause la règle générale et sans avoir recours à la notion d'exception. Cela passe par la reconnaissance d'une perte de mémoire de la langue, qui n'est pas sémantique — ce qui est un phénomène déjà bien établi — mais morphologique.

Il s'opère une «reconstruction» : contrairement à l'ensemble des autres mots en *-ion*, ces 88 mots ne peuvent pas être analysés comme dérivés d'une base verbale «préexistante». Ce sont au contraire les verbes qui sont reconstruits à partir des mots en *-ion* : sur la base thématique, extraite d'un mot en *-ion* régulièrement formé (base thématique), la langue reconstruit un verbe du premier groupe. Une fois entrée dans le lexique, la base empruntée s'impose pour tout le paradigme et peut donner lieu à la construction d'un verbe (*collecter*), d'un nom (*collectage*), etc.

Arrêtons-nous sur la série *collectage, collecter, collection,* particulièrement représentative de la situation. Cette série est en fait le résultat de découpages successifs dans le temps, qui aboutissent à une superposition de formations qui, du strict point de vue synchronique, pourrait sembler déroutante. Partant d'une base thématique *collec-t-*, empruntée au latin (supin *collectum*), la langue produit normalement *collec-t-ion*, sur base thématique et *collig-er,* sur base non thématique. Mais il arrive qu'une base cesse de fonctionner en tant que telle et ne soit plus reconnue. Ainsi, l'alternance *collec-t- / collig-* n'est plus perçue et fait place à la seule base *collec-t-*, disponible pour former de nouveaux mots : *collect-er* (signalé comme fautif dans le Larousse du XX^{e} siècle) et *collect-age* (d'apparition récente).

On retrouve bien ici la situation observée pour l'ensemble des couples *N-ion* et *V-er* sur bases thématiques et le jeu complexe de la mémoire de la langue[8].

On peut donc parler d'une perte de mémoire, dans la mesure où, lors de ces reconstructions, la langue «oublie» deux choses :

- l'existence, le cas échéant, d'une base non thématique (ou la possibilité d'en déduire une par comparaison avec d'autres familles de mots) et la possibilité de l'utiliser comme une base simple du français;

- le caractère thématique de la base utilisée : une fois le verbe créé, la base semble perdre son caractère thématique et se trouve alors disponible pour une dérivation avec un suffixe tel que *-age* (*collectage, détritage)*, qui sélectionne des bases non thématiques.

Les observations présentées ici ont pour conséquence que ces mots en *-ion* ne doivent pas être considérés comme exceptionnels : la règle générale qui décrit le processus de suffixation en *-ion* n'est pas remise en cause.

8 La même analyse rend compte de tous les *N-age* comportant la même base qu'un *N-ion* : *compactage, détectage, détritage, factage, fruitage, télédictage, visage.* Cf. Coret (1994), ainsi que d'autres verbes construits sur base thématique.

Mais au-delà de la simple explication du sous-ensemble de mots considéré (dont il faut bien de toutes façons rendre compte) cette analyse a mis en lumière un aspect particulier de la création lexicale — l'extraction de bases — et illustre aussi, en même temps que l'existence d'une mémoire des mots, son caractère lacunaire.

RÉFÉRENCES

CORET, M. (1994) : *Problèmes de suffixation et structuration du lexique,* Thèse de doctorat, Université Paris 7.

Dictionnaire Le Robert (1996) : version CD-ROM, Paris, Le Robert.

Le Grand Robert électronique (1989) : version CD-ROM, Paris, Le Robert.

HATZFELD, A., DARMESTETER, A. et A. THOMAS (1890-1900) : *Dictionnaire général de la langue française,* Paris, Delagrave.

HUOT, H. (1994) : «Sur la notion de racine», *TAL 35-2.*

HUOT, H. (1997) : «Des mots possibles aux mots existants : système morphologique et structuration du lexique», *Silexicales 1*, Lille, Université de Lille 3.

LITTRÉ, É. (1863) : *Dictionnaire de la langue française,* Paris, Gallimard - Hachette.

MERCIER (1801) : *Néologie ou Vocabulaire des mots nouveaux.*

PICOCHE, J. (1989) : *Dictionnaire étymologique du français,* Paris, Le Robert.

REY, A. (1992) : *Dictionnaire historique de la langue française,* Paris, Le Robert.

RICHELET, F. (1693) : *Dictionnaire françois contenant généralement tous les mots,* Genève.

ANNEXE : BASES THÉMATIQUES COMMUNES À *N-ION* ET *V-ER*

Le tableau ci-dessous regroupe la totalité des bases du français figurant à la fois, sans variation, dans un *N-ion* et un *V-er*. Pour chacune d'elle, j'indique l'étymologie latine (telle qu'elle est indiquée par Littré ou le *Petit Robert*), qui révèle bien, dans la grande majorité des cas, une opposition entre une base thématique et une base non thématique.

Dans les quatrième et cinquième colonnes figurent la base non thématique reconstruite ou attestée, disponible en théorie, et des mots révélateurs de la survivance de cette base non thématique. Les verbes possibles non attestés sont notés à l'aide du symbole °; ceux attestés en ancien français sont précédés de la mention «AF». La base non thématique pouvant résulter d'une formation populaire ou savante, on mentionne quelquefois deux mots. C'est le cas par exemple pour *collect-*, face à qui on a *collig-er* (formation savante) et *cueill-ir* (formation populaire).

	Base de *N-ion / V-er*	Etymologie latine du *N-ion*	Base non thématique	Mots attestés et verbes possibles (sur base non thématique)
1	***ablat***	ablatio, au-ferre	*-fer-*	cf. *conférer, déférer, souffrir* , issus de V latins en -ferre
2	***abras***	abrasio, ab radere	*-rad- / -rat-*	°*abrader* (cf. *radoire*, AF *ratoire*)
3	***accept***	acceptio, accipere	*-cip- / -cev-*	°*accevoir* (cf. *déception / décevoir*)
	except	exceptio, excipere		
	intercept	interceptio, intercipere, capere		
4	***affect***	affectio, afficere, facere	*-fai-*	AF *affaiter* ; °*affaire* (cf. *perfection / parfaire*)
	infect	infectio, infectus, inficere, facere		
	désinfect	mot préfixé, cf. infect-		
	réinfect	mot préfixé, cf. infect-		
5	***agress***	adgressio, adgredi, gradi	*-grad-*	cf *rétrograder, dégrader* issus de verbes latins en -gradi
	digress	digressio, digressum, digredi, gradi		
	progress	progressio, progredi, gradi		
	régress	regressio, regressus, regredi, gradi		
	transgress	transgressio, gradi		
6	***annex***	annexio, annectere		une seule base en latin
	désannex	mot préfixé, cf. annex-		
	connect/x	connexio, connectere		
	déconnect/x	mot préfixé, cf. connect-		
7	***assert***	assertio, asserere	*-ser-*	°*asserer* (cf *insertion / insérer*)
	désert	desertio, deserere		
8	***collect***	collectio, colligere	*-lig-*	*colliger ; cueillir*
9	***commun(i)***	communionem, communis		une seule base en latin
10	***compact***	compactus, compingere, pangere	*-ping-*	°*compinger ?*
11	***complét***	completus, complere, plere	*-pli-*	°*complir* (cf. *emplir, remplir* issus du latin -plere)
12	***complex***	complexionem, complexus, complectere		une seule base en latin
13	***compuls***	compulsio, compulsum, compellere	*-pell- / -pouss-*	AF *expeller, pousser* (cf *pousser* du latin -pulsus)
	expuls	expulsio, expellere		
	impuls	impulsio, impulsum, impellere, pellere		

	propuls	pro pulsum, pellere		
	répuls	repulsio, repulsum, repellere		
14	***concoct***	concoctio, coquere	*-cui(s)-*	°*concuire* (cf *cuire, décuire)*
15	***concrét***	concretio, concrescere	*-croît-*	°*concroître* (cf. *croître)*
16	***confess***	confessio, confiteri	*-fit-*	cf. *confiteor*
	profess	professio, professus, profiteri		
17	***contract***	contractio, contrahere, trahere	*-trai- / -tir-*	°*contraire* / °*contirer* (cf. *traction / traire ; tirer)*
	décontract	mot préfixé, cf. contract-		
	détract	detractio, trahere		
18	***détect***	detectum, detegere	*-tég-*	°*détéger* (cf. *protection / protéger)*
19	***détrit***	detritio, deterere	*-ter-*	°*détérer* ?
20	***dict***	dictio, dicere	*-di-*	*dire*
	édict	edictum		
21	***diffract***	diffractus, diffringere	*-fring-*	°*diffringer* (cf *diffringent, réfringent)*
	réfract	refractio, refringere		
22	***diffus***	diffusum, diffundere	*-fond-*	°*infondre* (cf *fusion / fondre)*
	infus	infusio, infusus, infundere		
	perfus	perfusio, perfusum, perfundere		
	transfus	transfusio, transfundere		
23	***dispers***	dispersio, dispergere, spargere	*-perg-*	°*disperger* (cf. *aspersion / asperger)*
24	***divis***	divisio, dividere	*-vid-*	°*divider* (cf. *dividende)*
	subdivis	mot préfixé, cf. divis-		
25	***édit***	editio, edere	*-di-*	°*édire* (cf *dire, dédire)*
	réédit	mot préfixé, cf édit-		
26	***éject***	ejectio, ejectum, ejicere, jacere	*-jic- / jet-*	AF *objicer, objeter , interjeter* (cf. *jeter)*
	inject	injectio, injicere		
	interject	interjectio, interjicere		
	introject	emprunté à l'allemand		
	object	objectio, objectum, objicere		
	réinject	mot préfixé, cf. inject-		

27	***électrocut***	emprunté à l'anglais	*-séqu- / -suiv-*	cf *consécution / conséquence ; suivre*
	exécut	exsecutio, exsequi		
	persécut	persecutio, per sequi		
28	***excis***	excisio, excisum, excidere, caedere	*-cid-*	*°excider* (cf. *décision / décider ; incident*)
	incis	incisio, incisum, incidere		
	précis	praecisio, praecisus, praecidere, caedere		
29	***excrét***	excretio, excretum, excernere, cernere	*-cern-*	cf *décerner, concerner ..* issus du latin -cernere
	sécrét	secretio, secernere, cernere		
30	***exempt***	exemptio, exemptum, eximere	*-im-*	*°eximer* (cf *péremption / périmer, rédimer*)
31	***explos***	explosio, explosum, explodere, plaudere	*-plod-*	*°exploder* (cf *applaudir* du latin applaudere)
	implos	formé sur explosion		
32	***inspect***	inspectio, inspicere	*-spic-*	cf *perspicace, auspice*
	prospect	prospectus, pro specere		
33	***invent***	inventio, inventum, invenire	*-ven*	*°invenir* (cf *prévention / prévenir*)
34	***invers***	inversio, inverteie	*-vert-*	*invertir*
35	***lés***	laesio, laesus, laedere	*-lid-*	cf *élision / élider*
36	***opin***	opinio, opinari		une seule base en latin
37	***opt***	optio, optare, opiare	*-op-*	*°opter ?*
	adopt	adoptio		
38	***préfix***	praefixus, praefigere	*-fig- / -fich-*	cf *ficher* du latin figere
39	***press***	pressio, primere	*-prim-*	*comprimer*
	compress	compressio, compressum, comprimere		
	décompress	mot préfixé, cf compress-		
40	***rébell***	rebellionem, rebellis		une seule base en latin
41	***recens***	recensionem, recensere		une seule base en latin
42	***révuls***	revulsio, revulsum, revellere, vellere	*-vell-*	*°réveller ?*
43	***transit***	transitio, transire	*-ir-*	*transir*
44	***vis***	visio, visum, videre	*-vid- / -voi-*	*voir*
	révis	visum, revisere		
	supervis	mot préfixé, cf. vis-		

L'*IMAGE* ET LA *FORME* : APLATISSEMENT OU DISTORSION DU TEMPS ?

Xavier LELUBRE

CRTT, Université Lumière Lyon 2, Lyon, France

1. LA TRACE DES ÉTATS ANTÉRIEURS DES CONNAISSANCES

L'image que nous avons d'une science et de sa terminologie, au moment présent qui est le nôtre, n'est-elle pas la forme figée qu'elle a alors prise ?

1.1 Chez Ibn al-Haytam, l'*image* et la *forme*

Le point de départ de cette recherche est l'existence de deux termes arabes d'optique employés de nos jours concurremment, qui correspondent au terme français (et aussi anglais) *image* : il s'agit de *sûra*[1] (c'est le terme le plus généralement employé dans le monde arabe, à commencer par l'Égypte) et *ḫayâl* (employé surtout en Syrie)[2].

Ce qui attire notre attention ici est le fait que ces deux termes, loin d'être d'utilisation récente en optique, se trouvaient déjà utilisés au Moyen Âge, mais où ils avaient deux acceptions différentes.

Dans la langue générale, *ḫayâl* a le sens de «fantôme, spectre, ce qui apparaît à un homme éveillé ou en songe»... et c'est aussi l'image que l'on voit dans un miroir[3]. Quant à *sûra*, c'est la forme (syn. *šakl*); extérieur, aspect, apparence; manière, façon (syn. *waẓh*); figure, image (représenté par la peinture ou le dessin).

---

[1] La transcription des caractères arabes utilisée ici est, dans l'ordre de l'alphabet oriental arabe habituel : */ ? b t ṯ ǧ ḥ ḫ d ḏ r z s š ṣ ḍ ṭ ẓ ` ġ f q k l m n h w y /.*

[2] L'Égypte et la Syrie ont été lors de la Renaissance arabe au siècle dernier (la *Nahḍa*) les grands pourvoyeurs de terminologie scientifique.

[3] Le grand dictionnaire *Lisân al-`Arab* d'Ibn Manzûr (m. 1311*)*, donne, entre autres : «*al-ḫayâl*^{u} *li-kull*^{i} *šay?*^{in} *tarâ-hu ka z-zill*^{i}, *wa-kaḏâlika ḫayâl*^{u} *l-?insân*^{i} *fî l-mir?ât*^{i} *wa-ḫayâl*^{u}*-hu fî l-manâm*^{i} *sûrat*^{u} *timṯâl*^{i}*-hi* » [ce qui apparaît de toute chose que l'on voit, comme l'ombre, de même, l'*image* de l'homme dans le miroir, et aussi : son *spectre* dans le sommeil est la forme sous laquelle il est représenté].

Nous pouvons voir la différence entre ces deux termes chez le célèbre savant arabe Ibn al-Haytam (m. vers 1039)[4], qui a joué au Moyen Âge un rôle fondamental dans l'évolution de l'optique, dans le passage suivant (cité par Mustafâ Nazîf 1942-1943 : 596)[5] :

> «L'image (*hayâl*) est la forme (*sûra*) de la chose vue (*mubsar*) que le regard (*basar*) saisit par la réflexion sur la surface [de séparation] du corps poli. Le lieu de cette image est le lieu où le regard saisit cette forme».

L'acception prise par ces lexies en optique ne peut être appréhendée que dans le cadre épistémologique où elles ont été établies comme termes. Le savant arabe établit une théorie de la lumière, une théorie de la vision[6], se démarquant dans son ouvrage, en particulier de la théorie antérieure des «rayons visuels», soutenue par différentes écoles depuis l'Antiquité.

1.2 Évolution des connaissances et terminologie

Comme le note Alain Rey (1979 : 64), avec l'évolution constante de leurs configurations conceptuelles, les sciences voient «leur terminologie évoluer, mais conserver forcément la trace des états antérieurs des connaissances : c'est donc un rapport changeant entre termes en partie anciens et notions nouvelles que la terminologie scientifique doit définir. Il s'agit ici avant tout d'une mise au point permanente des définitions.»

Ce sera bien le cas en optique, où des concepts physiques fondamentaux, comme ceux de lumière, de rayon lumineux, etc. ne cesseront d'être appréhendés, retravaillés, réinterprétés dans le cadre de théories concurrentes[7], tout en gardant les mêmes dénominations.

Ces dénominations pourraient-elles conserver une trace, en quelque sorte la mémoire de leurs acceptions antérieures ?

La mémoire est bien sûr liée au temps. Le déroulement du temps dans les civilisations n'est ni continu ni homogène. Il n'est qu'un lointain effet du temps physique exprimé par la succession des jours et des saisons et du temps exprimé par la succession des générations. C'est à cela que fait allusion le titre de cet article.

---

4 Connu en Occident sous le nom d'Alhazen, son ouvrage principal, *Kitâb al-Manâzir* (Le livre de l'optique) sera traduit en latin (par Risner, en 1572), sous le nom d'*Opticae Thesaurus Alhazeni Arabis*, ouvrage qui eu une influence considérable dans ce domaine en Europe.

5 Mustafâ Nazîf (1942-1943 : 103) indique comment, chez cet auteur, *sûra* a le sens du terme chez les philosophes (la *forme*, opposée à la *substance* (*hayûlâ*) et *hayâl*, c'est l'*image* dans l'acception moderne. Voir aussi, sur *sûra* (chez Ibn al-Haytam) traduit par *forme*, Roshdi Rashed (1970 : 278-280).

6 Sur l'optique chez Ibn al-Haytam, voir en particulier Gérard Simon (1989), Roshdi Rashed (1993).

7 Ibn al-Haytam développe une théorie de la lumière. Bien plus tard viendront les théories électromagnétiques, corpusculaires, quantiques sur la nature de la lumière, toujours dénommée, au-delà de ces approches, *lumière*. D'un autre point de vue, sur les termes latins, de *lumen* et *lux*, pour la lumière, voir Vasco Ronchi, 1966 : 103.

Le temps ici considéré est sa matérialisation dans l'évolution des composantes d'une civilisation donnée, et pour ce qui nous intéresse plus particulièrement, parmi les domaines de la connaissance scientifique et de la pratique technologique, celui de l'optique.

L'évolution de ces différents domaines est généralement discontinue et différentielle, chacun évoluant à sa vitesse, avec des pauses, des accélérations, voire au contraire des retours en arrière. Il y a forcément distorsion entre le *temps* d'un domaine donné et le *temps* de la société.

Cette évolution est notamment liée à celle des autres domaines — ainsi l'optique et les mathématiques, l'astronomie, la fabrication des composants optiques, le développement d'autres branches de la physique, comme la mécanique, l'électromagnétique...—, à l'état général de la société considérée, et à ses relations avec d'autres civilisations, directement ou indirectement — par les traductions — l'optique en Occident s'est développée au Moyen Âge en grande partie sur la base des traductions latines de documents en arabe, eux-mêmes redevables aux traductions en syriaque puis en arabe de documents grecs de l'Antiquité[8].

Les connaissances, les théorisations, les concepts, les savoir-faire, les procédés, les outils constitutifs de ces domaines — les *notions* ou *concepts* de la terminologie, que nous appelons *unités référentielles* — s'expriment par les terminologies qui leur sont associées.

Les unités référentielles d'un domaine sont liées à un état donné de la constitution de ce domaine. Leur histoire relève de l'histoire des civilisations, des idées, des sciences et techniques. Quant aux terminologies qui leur sont associées, elles constituent le matériau de cette étude.

2. L'OPÉRATION DE DÉNOMINATION ET LA CAPTATION D'EFFETS DE MÉMOIRE

L'unité référentielle d'un domaine donné, établie par ceux qui l'ont «découverte», l'est avec sa dénomination. Celle-ci peut se trouver modifiée par d'autres instances — et elle l'est bien évidemment quand elle a à être dénommée dans d'autres langues que celle de ceux qui ont établi l'unité référentielle.

La question se pose de savoir si lors de l'opération de dénomination d'une unité référentielle, le mécanisme lui-même de cette opération ne conduirait pas à capter des éléments extra-terminologiques, qui correspondraient à une époque et à un moment de l'évolution de la spécialité concernée, trace d'événements qui y sont reliés.

Ces traces pourraient constituer la *mémoire* d'un terme.

La mémoire des termes serait-elle alors l'un des constituants de la partie inavouable, soigneusement évacuée par la terminologie, leur connotation ? C'est-à-dire le lien d'un terme avec les autres mots de la langue, au-delà des rapports de ce terme avec les autres termes du «système terminologique» — appellation d'ailleurs impropre[9] — dont il ressort ?

---

[8] Parmi une abondante bibliographie sur l'optique dans le Monde islamique, Bernard Maitte (1981 : 21-26), A. I. Sabra, «Manâzir ou `Ilm al-Manâzir», in *Encyclopédie de l'Islam*, tome IV (2e éd.).

[9] Ce qui fait système, ce sont les unités référentielles du domaine ou du sous-domaine référentiel, et non pas, généralement, les dénominations. Ainsi, en optique géométrique, ce

2.1 Le cas de la terminologie arabe

Si nous considérons l'état actuel de la terminologie arabe de l'optique, celle-ci a un fonds ancien auquel nous avons fait allusion précédemment. Mais le plus gros de cette terminologie est bien plus récent. En particulier, on sait, en ce qui concerne les terminologies scientifiques et techniques arabes contemporaines l'importance, pour des raisons historiques, du français et de l'anglais. C'est essentiellement à travers ces deux langues — et il faut tenir compte de ce que chacune est ou a été la langue de référence utilisée dans telle ou telle région du Monde arabe bien souvent à l'exclusion de l'autre —, que le Monde arabe contemporain participe au mouvement scientifique mondial.

Le premier contact avec les terminologies scientifiques arabes peut déconcerter le traducteur, tant est importante la question des phénomènes de variation terminologique d'une région arabe à une autre, voire d'un auteur à un autre.

La profusion de termes synonymes — que les instances interarabes s'attachent à limiter — est par elle-même signifiante. Elle ressort d'au moins trois faits :

- l'émiettement de la création terminologique, qui peut être considéré comme une conséquence du manque d'unité politique;

- l'influence de deux centres arabes créateurs de terminologie, l'Égypte et la Syrie;

- l'existence de deux sources étrangères de terminologie, la source française et la source anglaise.

C'est ce que nous disent, comme nous allons le voir, de nombreux termes de l'optique, tant par l'existence de séries de termes synonymes, que, par exemple, de faits de polyvalence ou d'hypéronymie.

2.2 L'impondérable part de mémoire captée par les différents types de dénomination

L'opération de dénomination d'une unité référentielle consiste, si le terme est motivé, à exprimer un sous-ensemble[10] des traits de substance, des caractéristiques de cette unité référentielle. Ce choix n'est pas livré au hasard.

Deux situations sont possibles : celle où la dénomination est faite dans la langue de celui ou ceux qui ont établi cette unité référentielle et celle où la dénomination est reprise dans d'autres langues[11]. Dans ce dernier cas, deux possibilités existent : exprimer cette dénomination dans la langue d'accueil — au moyen des différents procédés possibles offerts par cette langue — ou bien procéder à une dénomination autonome, à partir de l'unité référentielle elle-même, par choix éventuellement d'autres traits de substances. Dans

---

sont l'objet, l'image, l'espace objet, l'espace image, le système optique, etc., qui font système, et non pas les termes *objet*, *image*,... Les procédés les plus divers de création de termes, la diachronie dans laquelle cette création s'inscrit; s'opposent à ce qu'ils constituent un système, sauf pour des sous-ensembles bien limités.

10 La dénomination est une opération extrêmement réductrice. Elle ne peut exprimer que quelques traits de substances de l'unité référentielle.

11 Ce sont respectivement la *néonymie d'origine* et la *néonymie d'appoint* de Guy Rondeau (1983 : 124)

le premier cas, l'on s'appuie avant tout sur la dénomination de la langue source; auquel cas, il y a forcément, d'une façon ou d'une autre, une opération de calque : la dénomination est guidée par l'original.

En particulier, le cas du calque, qui peut fonctionner dans les deux sens. Ainsi au Moyen Âge, de l'arabe vers le latin, puis le français : la *rétine* : *sabakiyya*, l'*humeur cristalline* : *rutûba zalîdiyya*, la *camera obscura* (la chambre obscure, aujourd'hui encore, c'est le terme latin, du Moyen Âge, qui est généralement utilisé en français), *bayt muzlim*.

L'influence de la langue d'origine peut s'exprimer aussi de bien d'autres manières, au niveau de la synonymie, à celui de la polyvalence ou de l'hyperonymie.

Ainsi, pour la dénomination d'instruments à vision éloignée, le français appelle généralement *télescope* l'instrument dont l'objectif est un système réflecteur et *lunette* l'instrument dont l'objectif est un système réfracteur[12]. Les équivalents anglais courants sont, respectivement, *reflecting telescope* et *refracting telescope*. L'anglais possède de ce fait un terme hyperonyme, *telescope*, tandis que le français... n'en possède pas[13]. Ces divergences terminologiques se réfractent, si l'on peut dire, dans les terminologies arabes, selon qu'elles suivent le français ou l'anglais.

À nouveau deux cas se présentent : ou la dénomination choisie est une lexie qui existe déjà dans la langue, mais à laquelle on donne une acception nouvelle, pour le domaine de spécialité considéré, ou bien c'est une lexie nouvelle qui est forgée. Dans ce dernier cas, elle utilisera des éléments de la langue — morphèmes ou lexies constituantes — existant déjà. Dans les deux cas, de toute façon, la nouvelle unité terminologique ne sera jamais sans lien avec d'autres. Bien sûr, ces liens diffèrent d'une langue à une autre.

La langue arabe dispose de quatre procédés pour créer ses termes : le recours à son système de nomination[14] (les procédés morphologiques), le recours à son système de communication (les procédés syntaxiques), le recours aux transferts sémantiques et, enfin, l'emprunt.

a) le recours au système de nomination

Il aboutit à la formation d'unités terminologiques simples. Nous avons vu le cas de *hayâl*, *image*. La lexie est ancienne, attestée par la poésie, elle a plusieurs emplois et, en particulier, celui d'image vue dans un miroir. L'arabe ancien de spécialité a tout naturellement repris cette lexie, en en précisant l'acception dans le domaine de l'optique. Il en est ainsi pour de très nombreux termes.

---

12 Les télescopes s'appellent aussi *réflecteurs* ou encore *télescopes catoptriques*. Les lunettes portent aussi le nom de *réfracteurs* ou *télescopes dioptriques* (H. Pariselle, *Les instruments d'optique*, Paris, Armand Collin, 1933 : 99)

13 On utilise alors l'expression «les lunettes et les télescopes».

14 Sur les systèmes de nomination et de communication de l'arabe, voir André Roman (1989).

En revanche, le terme *istiqtâb*, l'équivalent arabe de *polarisation*[15], n'est pas attesté dans les dictionnaires arabes anciens. Il a été créé très probablement en référence à la lexie *qutb*, dont l'un des sens est *pôle*, de racine (*q* - ***t*** - *b*), comme nom d'action du verbe dérivé *istaqtaba*. Cette racine, elle, est ancienne et a plusieurs sens, qui nous importent peu ici. Mais on ne peut évacuer, pour le locuteur arabophone, les résonances de sens que peut induire la racine, ici triconsonantique, mais qui peut être quadriconsantique[16].

Il existe ainsi pour chaque terme arabe formé dans le cadre du système de nomination, c'est-à-dire, formé sur une racine et un schème, des relations inévitables avec les autres lexies ayant la même racine, ou voire le même schème[17].

b) le recours au système de communication

Ce procédé aboutit à la formation d'unités terminologiques complexes, dont les éléments sont ou ne sont pas des termes.

Ainsi, à titre d'exemples, les termes *istiqtâb dâ ?iriyy yamîniyy*, *polarisation circulaire droite*, *mizhar tabâyun* ***at-tawr***, *microscope à contraste de phase*.

Ce mode de formation est fréquemment sollicité en arabe. Celui-ci ne dispose pas, pour la formation de sa terminologie scientifique, d'une langue de prestige, comme le grec et le latin pour le français, où il pourrait puiser pour créer des doublets ou des formants. De nombreux termes arabes formés par recours au système de communication sont alors des équivalents de termes français (et anglais) qui le sont grâce à la composition savante, procédé très peu fréquent en arabe. Ils en sont le développement syntaxique.

Par exemple, l'un des équivalents arabes de *photoélectricité* est *kahrabâ ? **d**aw ?iyya* («électricité lumineuse»)[18,19,20].

---

15 De plus, dans cet exemple, il y a trois termes *polarisation* : l'un concerne l'électromagnétisme, l'autre l'électrocinétique et enfin, le troisième, l'optique. Le terme arabe n'a fait ici que suivre les tribulations du terme français (ou anglais).

16 Pour ce qui concerne les termes eux-mêmes, nous savons, par exemple, comment les instances terminologiques veillent à éviter les mots tabous. Dans le domaine arabophone, ils peuvent de plus varier d'une région à une autre.

17 Ce type de connotation est du domaine de la psycholinguistique et devrait faire l'objet de tests.

18 La photoélectricité est le phénomène électrique d'origine lumineuse. L'on trouve aussi, comme équivalent, le terme de ***d**aw? kahrabâ?iyy* («lumière électrique»), ce qui n'est pas la même chose.

19 Le terme *kahrabâ?* vient du persan *kahrobâ* («ambre jaune, succint»). C'est un emprunt ancien. L'on peut comparer cela au terme français *électricité*, formé sur le grec *êlektron* («ambre jaune, succint» - *Dictionnaire Grec-Français* de Bailly). Quant aux Persans, pour *électricité*, ils utilisent le terme *barq*, emprunté... à l'arabe («éclair»; le télégramme se dit en arabe *barqiyya*). On a là l'exemple d'un bel échange terminologique ! Selon Vincent Monteil (1960 : 33;134), c'est Rifâ`a Râfi` at-Tahtawî (1801-1873) qui a le premier utilisé ce terme, en 1834, dans sa relation de voyage en France, *Talhîs al-Ibrîz ?ilâ talhîs Bârîz* (voir la traduction française *L'or de Paris.- Relation de voyage, 1826-1831,* par Anouar Louca, Paris, Sinbad, 1988 : 123).

20 D'autres synonymes existent, qui font recours à des formants créés à cet effet, tels que ***d**awkahrabâ?iyya* et... *kahra**d**aw?iyya*.

Si, en français ou en anglais, la facture de lexies à l'aide de formants grecs ou latin, indique d'emblée que ce sont — très probablement — des termes scientifiques, le développement de ces éléments dans le cadre de la syntaxe arabe, avec des éléments qui ne sont pas forcément des termes, est bien moins parlant. Ainsi *qarîb min al-mi**h**war* «proche de l'axe», équivalent arabe de *paraxial* est beaucoup moins marqué de ce point de vue[21] que le terme français, ou que son synonyme *bârâmi**h**wariyy*, créé, lui, avec le formant emprunté *bârâ-*.

Les unités terminologiques complexes, plus longues que les unités simples, sont susceptibles de contenir davantage d'information, d'être plus motivées que celles-ci. Mais cela n'est qu'apparence, n'enlevant rien au caractère réducteur de la dénomination. Ainsi l'exemple ci-dessus de *kahrabâ ? **d**aw ?iyya* ne fait sens que grâce à la connaissance du référent.

c) le recours aux transferts sémantiques

Nous avons évoqué plus haut le fréquent réemploi de lexies existant déjà. Nombreux sont les termes d'optique formés par recours à la métonymie, à la métaphore et à l'hypallage. La plupart du temps l'on retrouve les mêmes glissements de sens qu'en français ou en anglais[22].

Ainsi, en ce qui concerne la métaphore, l'équivalent arabe de *lentille* (*'lens'* en anglais pour la pièce optique, et *'lentil'* pour le légume) est *`adasa* : ce terme est inconnu en optique chez les savants arabes du Moyen Âge[23]. Il s'agit bien du légume bien connu. Ce terme a en Europe une histoire intéressante (Vasco Ronchi, 1966 : 29-30) : la découverte des verres de lunettes «se produisit entre 1280 et 1285, très certainement dans la vallée de l'Arno», probablement par quelque artisan vitrier, par hasard, qui permettaient de corriger la presbytie :

> «À cette époque, les scientifiques, bien que connaissant cette affection (le nom en est grec et très ancien), ne savaient pas du tout à quoi elle était due; non plus d'ailleurs que l'effet optique des disques de verre à faces convexes. Ces disques furent appelés *lentilles de verre* à cause de l'analogie de forme qu'ils offraient avec les lentilles comestibles. Un tel nom est une autre preuve de l'origine artisanale de cette invention. Jamais, à l'époque, un scientifique n'aurait donné le nom d'un légume à l'une de ses découvertes».

---

21 Dans ce syntagme, seul l'élément *mi**h**war* (*axe*) est lui-même un terme. Un autre équivalent arabe existe aussi pour *paraxial*, c'est *mutama**h**wir*, formé dans le cadre du système de nomination sur la racine quadriconsonantique (m - **h** - w - r), elle-même créée à partir du terme *mi**h**war*, lui-même de racine triconsonantique (**h** - w - r) : c'est un exemple de création de racine à partir d'un terme.

22 Voir sur cette question Lelubre (1992, chapitre 3.3).

23 Ibn al-Haytam, parmi d'autres savants, avait fait des expériences, des montages où des lentilles étaient utilisées, mais il n'a pas de termes pour les désigner. Cependant, Roshdi Rashed (1993 : 233-234), à propos du terme *ballûr* ou *billûr*, cristal de roche, emprunt, avec métathèse, du grec *bêrullos*, évoque un passage du minéralogiste at-Tîfâsî (m. 1253) et un autre, d'un astronome, Taqî ad-Dîn b. Ma`rûf (ouvrage achevé en 1574), où le terme *ballûra* est utilisé, qui semble indiquer une lentille plan-convexe, en cristal de roche.

Quant aux lentilles à faces concaves, comme «aucune à faces creuses n'existait parmi les légumes, les verres pour myopes ne furent pas nommés *lentilles* mais *verres creux* ».

Pour ce qui est de la métonymie, nous avons, par exemple, le cas de ***ad-daw*** *?*, la *lumière*, employé pour *al-basariyyât, l'optique*, comme en français, et de même *su`â` **daw** ?iyy, rayon lumineux* employé comme *su`â` basariyy, rayon optique.*

Le recours aux éponymes représente un cas particulier et fréquent de métonymie; les éponymes constituent un lieu de mémoire flagrant. Ainsi nos bien françaises *lois de Descartes*[24], ailleurs *lois de Snell* (ou de *Snellius*), ou encore, plus oecuméniques, *lois de Snell-Descartes* deviennent en arabe, selon que prédomine l'influence française ou l'influence anglo-saxonne — c'est-à-dire que sont adoptés les points de vue français ou anglo-saxons —, *qânûn-â Dîkârt, qânûn-â Snîl*, et dans certains ouvrages, elles deviennent carrément arabes : c'est *qânûn-â Ibn al-Haytam*, les *lois d'Ibn al-Haytam*[25] !

Comme exemple d'hypallage, le *réseau sinusoïdal, **muh**azzaza zaybiyya*, n'est pas un réseau de diffraction dont la forme serait sinusoïdale, mais en fait un réseau dont la périodicité est sinusoïdale.

d) l'emprunt

L'emprunt, bien sûr, indique la provenance étrangère. Il peut concerner une unité terminologique entière ou il peut être seulement celui d'un formant. Il peut aussi d'une part permettre de différencier des sources étrangères et d'autre part de donner des indications sur les procédés de la naturalisation du terme dans la langue d'accueil. Prenons deux exemples en arabe.

L'emprunt nous en dit aussi encore davantage, si l'on considère la façon dont il se fait sur le simple plan phonétique. Ainsi à côté de *mîkrûskûb* (sur la prononciation française), l'on trouve, *mâykrûskûb* (prononciation anglaise).

Évocatrice est aussi la transcription de certains noms de savants, qui n'étaient ni francophones ni anglophones. Ainsi *Kirchhoff* (1824-1887), allemand et germanophone, devient en arabe soit *Kîrsûf* soit *Kîrtsûf*, selon la prononciation française ou anglaise. Ces transcriptions ne sont-elles pas révélatrices d'une vision, induite par l'histoire contemporaine, francophone ou anglophone du reste du monde ?

---

24 On disposait auparavant de tables de réfraction, grâce notamment à Witelo. Selon Bernard Maitte (1981 : 70), il semblerait que Descartes (sa *Dioptrique* date de 1637) ait eu directement connaissance des travaux de Snell, décédé prématurément, qui aurait trouvé les relations exactes entre angles d'incidence et de réfraction vers 1625.

25 La loi de la réflexion (égalité de l'angle d'incidence et de l'angle de réflexion) ne posait pas problème. Il n'en était pas de même pour la réfraction, pour laquelle si la mise au point de tables précises de réfraction a été faite chez les Arabes — Ibn al-Haytam en a établies —, l'établissement d'une relation mathématique était autre chose.

La graphie arabe utilisée[26] peut être parlante : ainsi le phonème non arabe /g/, que toutes les institutions interarabes recommandent de transcrire par le graphème arabe *gayn*, est généralement rendu par le graphème *zîm* en Égypte (où il est prononcé [g]; le phonème /z̲/, quant à lui, est rendu par un graphème de même forme, mais comportant trois points au-dessous au lieu d'un seul; mais ce graphème, avec trois points dessous est utilisé en Irak, à l'instar des Iraniens, pour représenter le phonème /ts̲/), par le graphème *kâf*, surmonté d'un trait oblique (à l'iranienne, en Irak) ou de trois points (au Maroc), ou encore, comme souvent en Tunisie, par le graphème *qâf*, surmonté ou non de trois points[27], diacritisation qui à elle seule, évoque une différence entre le Moyen-Orient et le Maghreb sur le plan calligraphique[28].

L'emprunt peut s'effectuer selon différents degrés d'intégration. Ainsi pour le terme *fluorescence*, on trouve en arabe deux équivalents : *fulûriyya* et *tafalwur*. Le premier est formé par l'adjonction du suffixe *-iyya* sur l'emprunt *fulûr*, ou *filûr*, voire *fallûr* ou *falwar*; ce suffixe a de nombreux emplois, dont celui qui correspond aux suffixes *-escence*, *-ité*, *-isme* du français. Le second, quant à lui, est aussi formé à partir de l'emprunt *fluor*, mais de manière bien différente : de l'emprunt a été extraite une racine quadriconsonantique, la racine (*f - l -w - r*), sur laquelle ce terme a été formé, grâce au système de nomination de l'arabe, selon le schème [*ta*R_1*a*R_2R_3*u*R_4], R_1 étant la première consonne radicale, etc. De ce fait c'est un terme de facture arabe, formé sur une racine, elle, créée à partir d'un emprunt[29].

Des différents exemples évoqués ci-dessus, il apparaît que l'ont peut classer les termes en deux catégories : ceux pour lesquels la forme nous donne des indications qui dépassent le cadre de la terminologie (présence d'éponymes, emprunts, formations savantes pour les termes français et anglais; présence de formants ou d'affixes spécifiques), et les autres, dont la forme — morphologie, syntaxe — ne les distingue en rien des lexies de la langue commune[30]. Pour le premier groupe les termes présentent de manière intrinsèque des signes extra-terminologiques. Ce n'est qu'une approche extérieure qui peut nous renseigner sur des faits extra-terminologiques concernant les termes de l'autre groupe.

---

26 En l'absence de convention véritablement respectée dans tous les pays arabes de transcription en caractères arabes de phonèmes non arabes.

27 Ou tout simplement le graphème *qâf*, prononcé [g] dans de nombreux dialectes, dans le Monde arabe.

28 Il s'agit de la différence de diacritisation des graphèmes *qâf* et *fâ?*.

29 Ce procédé est très ancien et explique la création de bon nombre de racines quadriconsonantiques en arabe. Les racines créées de nos jours, pour des besoins terminologiques, sont majoritairement de ce type.

30 Par exemple il n'existe pas en arabe de doublet savant d'une lexie courante, qui serait dû à une évolution phonétique différente (comme cela arrive en français), ou de doublets de racines consonantiques (comme c'est le cas pour l'hébreu, qui peut faire recours à des racines araméennes). En revanche, les racines arabes formées à partir de lexies arabes ou d'emprunts, elles, sont porteuses de nouveauté par rapport au stock des racines de la langue.

3. LA CHARGE DE MÉMOIRE D'UN TERME

3.1 Les types de relations qu'un terme peut avoir

L'on peut se demander quels sont les types de relations (au sens le plus large) qu'un terme peut entretenir avec certains types d'entité, relations susceptibles d'être porteuses de mémoire. Deux types s'imposent :

(a) les relations référentiellement induites

- le terme avec l'unité référentielle qu'il dénomme (en particulier sa motivation);

- le terme avec les autres termes du «système terminologique» dont il dépend;

- le terme dans des rapports d'hyperonymie, comme composant d'autres termes, dans des rapports de synonymie;

- le terme utilisé dans un autre domaine;

- le terme déterminologisé, entrant dans la langue commune.

Ces relations peuvent varier avec l'évolution du domaine.

(b) les relations de type lexical

- la facture du terme (ses composants, ses formants, la façon dont il a été formé, les phénomènes de calque,...);

- avec d'autres lexies de la langue;

- la connotation du terme.

3.2. Conclusion : aplatissement et distorsion du «temps»

Dans bien des domaines, la terminologie apparaît comme partiellement incohérente; cela est vrai si on la considère d'un point de vue strictement synchronique, si l'on fait abstraction de son histoire, de sa construction au cours du temps : il y a alors, dans cette façon de considérer les faits, ce que l'on pourrait appeler un *aplatissement* du «temps». Le temps, pour les terminologies, nous l'avons vu, c'est celui qui commande l'évolution, c'est-à-dire la construction, l'organisation, la réorganisation en fonction de nouveaux éléments de chacun des domaines concernés — comme nous l'avons évoqué pour l'optique —, évolution non linéaire.

Une telle vue strictement synchronique élimine, bien entendu, tous les termes qui ont disparu, soit en raison de la réorganisation du domaine — disparition d'unités référentielles — ainsi, l'*éther* du siècle dernier, depuis l'expérience de Michelson et Morley —, soit en raison de changements, de rectifications terminologiques, comme par exemple, dans la terminologie syrienne de la physique, pour l'*énergie*, le terme *qudra*, remplacé par le terme *tâqa*, utilisé partout ailleurs.

Mais, dans le cas de l'arabe en particulier, où coexistent, synchroniquement des variantes terminologiques, la prise en compte des variations régionales est par contre inévitable.

Ce que nous pouvons appeler *distorsion* du «temps» correspond à l'évolution différentielle au cours du temps d'un domaine par rapport à un autre, avec des phénomènes d'inertie terminologique, d'où les phénomènes de distorsion entre ou au sein de terminologies.

RÉFÉRENCES

TERMINOLOGIE, LINGUISTIQUE

REY, Alain (1979) : *La terminologie : noms et notions*, Paris, PUF, 1re éd., 128 p.

RONDEAU, Guy (1993) : *Introduction à la terminologie*, Québec, Gaëtan Morin, 2e éd., 238 p.

ARABE

LELUBRE, Xavier (1992) : *La terminologie arabe contemporaine de l'optique : faits - théories - évaluation*, Thèse de Nouveau Doctorat, Université Lyon 2, 546 p.

MONTEIL, Vincent (1960) : *L'arabe moderne*, Paris, Klincksieck, 386 p.

ROMAN, André (1990) : *La grammaire de l'arabe*, Paris, PUF, coll. «Que sais-je ?», 128 p.

OPTIQUE

MAITTE, Bernard (1981) : *La lumière*, Paris, Seuil, coll. «Points», 345 p.

NAZÎF, Mustafâ (1942-1943) : *al-**H**asan b. al-Haytam : buhûtu-hu wa kusûfu-hu* [*// Ibn al-Haytam : recherches et découvertes //*], 2 volumes, Le Caire, Matba`a Nûrî, 879 p.

RASHED, Roshdi (1970) : «Optique géométrique et Doctrine optique chez Ibn al-Haytham», *Archive for History of Exact Sciences*, vol. 6, n°4, pp. 271-298.

RASHED, Roshdi (1993) : *Géométrie et dioptrique au Xe siècle — Ibn Sahl, al-Qûhî et Ibn al-Haytham*, Paris, Les Belles Lettres, 315 + 7 p.

RONCHI, Vasco (1966) : *L'optique, science de la vision*, Paris, Masson, 158 p.

SIMON, Gérard (1989) : «L'optique d'Alhazen et la tradition ptoléméenne : une nouvelle insertion dans le champ du savoir», *Colloque International d'histoire des sciences et de la philosophie arabes*, Paris, Institut du Monde Arabe, 22-25/11/1989, 7 p.

DE L'EMPLOI LIBRE À L'EMPLOI SUPPORT

Hassane FILALI SADKI

Université de Franche-Comté, Besançon, France

Le processus qui conduit un élément lexical — verbe, nom, préposition,... — à devenir un support de prédication reste encore un mécanisme mal connu et pas assez étudié. Nous essayerons de voir comment s'effectue le processus de délexicalisation d'une entité donnée qui perd soit totalement, soit partiellement sa valeur pleine pour ne devenir qu'un élément presque transparent à la relation prédicative. Toutefois, on constarera que dans cette perte significative de la valeur lexicale de base, l'élément qui devient support de prédication d'un N prédicatif garde une emprunte de sa fonction lexicale originale.

Notre hypothèse est que l'élément lexical, candidat à la fonction support, se vide dans son parcours de sa valeur élémentaire. Il a été considéré, jusqu'à présent, dans les différents travaux, que c'est le verbe qui sélectionnait les substantifs avec lesquels il se combinait. Or, les travaux récents, menés autour de M. Gross au LADL et G. Gross au LLI, démontrent plutôt l'inverse. Il s'est avéré que c'est les N prédicatifs qui sélectionnaient les verbes supports. Ces derniers n'ont comme fonction, en plus de celle d'être supports de la prédication et actualisateurs du temps, du nombre et de la personne, d'exprimer une certaine modalité d'action et l'aspect. Cette modalité d'action ne leur est pas inhérente dès la base lorsqu'ils sont employés dans leur valeur lexicale pleine d'une part et que les verbes qui ont cette particularité constituent une classe très réduite. Chaque item prend alors une valeur «modalisatrice» particulière en fonction du nom prédicatif qui le sélectionne.

Nous verrons par la suite que l'une des particularités des supports est qu'ils télescopent et figent une partie de leur construction, de sorte que toute modification de la structure syntaxique devient fatale à la grammaticalité de la phrase. Ainsi, si l'on prend pour exemple le verbe *prendre*, nous constatons qu'il y a une distorsion entre sa valeur lexicale pleine en tant que verbe d'action exprimant un processus et sa valeur délexicalisée où il est support.

L'étude des emplois spécifiques de *prendre*, en tant que verbe support ou opérateur, n'est possible qu'en faisant une analyse de toutes les structures syntaxiques dans lesquelles il s'insère, et l'analyse de tous les substantifs qui peuvent lui être associés. Notre étude portera essentiellement sur les emplois libres, les emplois supports et en fin les emplois figés. Nous nous interrogeons sur les mécanismes qui permettent à la même entité lexicale

d'avoir des emplois aussi variés et diversifiés et comment on peut trouver un lien entre eux tout en justifiant les changements à la fois syntaxiques, sémantiques et fonctionnels. Ainsi, pour éviter des confusions, il est nécessaire de spécifier les propriétés de chaque construction. Cette nécessité d'éclaircissement s'impose dans la mesure où le lexique n'est pas monosémique et que les verbes supports ont, le plus souvent, des valeurs différentes.

EMPLOIS INTRANSITIFS ET PRÉPOSITIONNELS

- (le vaccin + le feu + cette plante + le mortier + le plâtre) a bien pris

Ce type de phrase présente des propriétés syntaxiques particulières, et leur constructions diffèrent de celles que nous analysons pour *prendre*. Tout d'abord, ces constructions en *prendre* n'ont pas de compléments prédicatifs, ce qui est l'objet de cette étude. D'autre part, pour certains types de ces constructions, il est possible d'avoir une construction factitive.

- Luc fait prendre (le feu + la mayonnaise)
- (le feu + la mayonnaise) prend

Par contre, des constructions similaires à celles-ci font l'objet de notre étude.

- la teinture prend dans ce tissu
- ce tissu prend la teinture

Il en est de même de :

- Le feu prend dans cette maison
- cette maison prend feu

On peut dire que, dans la perspective du «lexique-grammaire», les entrées du dictionnaire ne sont pas des mots isolés, mais plutôt des phrases. Par conséquent, il y a autant de verbes, pour une unité définie morphologiquement comme telle, que de constructions transformationnellement possibles

EMPLOIS CONCRETS

Les emplois concrets de *prendre* acceptent la structure à trois arguments N_0 prendre N prép N_1, Les arguments de *prendre* peuvent être isolables à l'aide de questions :

1- Luc a pris un livre à Max
- qui a pris un livre à Max ?
- Luc
- qu'a pris Luc à Max ?
- un livre
- à qui Luc a-t-il pris un livre ?
- à Max

Ce test de questionnement permet de spécifier que $N_0 = N_{hum}$, N_1 = N concret et $N_2 = N_{hum}$. Notons que le complément N_2 n'est pas nécessairement un N_{hum}, comme dans le cas de l'exemple suivant:

2- Luc a pris un livre à la bibliothèque

Dans ce cas, la question appropriée est où :

- où Luc a pris le Livre ?
- à la bibliothèque

Lorsque le verbe *prendre* a un complément concret, il a un emploi libre, et de ce fait il est possible de lui substituer d'autres verbes.

- Luc a (emporté + emprunté + volé + feuilleté) ce livre

Dans cette structure, le verbe *prendre* exprime un changement de localisation dans l'espace, puisqu'il y a un avant et un après. Ainsi, pour (1) et (2) on peut dire que avant : Max avait le livre ou qu'il (était + se trouvait) à la bibliothèque et qu'après : Max n'a plus le livre ou que le livre n'est plus à la bibliothèque

PRÉSENTATIONS DES LISTES

Les substantifs représentés dans les listes sont des prédicats nominaux qui forment avec le verbe support *prendre*, dans une phrase simple, une entrée lexicale autonome. Ces N prédicatifs sont répartis dans différentes tables en fonction de la nature du substantif et de celle du déterminant. Ainsi nous avons soit des N libres (à déterminants non contraints) soient des N non libres (à déterminants figés). À l'intérieur de cette bipartition, les N prédicatifs sont classés d'après leurs constructions syntaxiques.

Dans les listes à déterminant figé, on a remplacé le symbole N par le symbole C pour indiquer la présence d'un élément figé. Étant donné que le déterminant est figé, on peut dire qu'il fait partie du prédicat nominal. C'est la raison pour laquelle, l'élément C représente à la fois le N prédicatif et son déterminant. Ces constructions figées sont réparties en tables en fonction de leur structure syntaxique. À l'intérieur de chaque structure, nous avons établi des sous-structures en fonction de la nature de l'élément nominal, du modifieur : adjectival ou prépositionnel.

PROBLÈMES GÉNÉRAUX DE SYNTAXE ET DE LEXIQUE

Les prédicats nominaux que nous étudions sont classés en fonction de leur construction, et plus particulièrement en fonction de la nature de leur argument. La première démarche consiste à distinguer d'abord les N prédicatifs qui prennent un complément de ceux qui sont en construction intransitive. Les compléments des V-n sont généralement de forme prép N_1. Ainsi, en fonction de ce critère on peut avoir les différentes structures où les N prédicatifs ont pour support le verbe *prendre*.

(1) N_0 prendre (E + Dét) N
- Luc prend l'air

- la maison prend feu

(2) N_0 prendre (dét N) (prép N_1)
- Luc prend Max sous sa direction

(3) N_0 prendre (Dét N) (à N_1)
- Luc prend cette proposition au sérieux

(4) N_0 prendre (dét N) (de N_1)
- Luc prend des informations de Max

(5) N_0 prendre N_1 (prép N)
- Luc prend Max en charge
- Luc prend Max en haine

Donc, le critère de la complémentation nous permet d'établir une classification provisoire, et d'avoir trois types de constructions, celles présentées en (1) sans complément, la seconde comprend un complément prépositionnel, les cas de (2) à (4). Ces trois structures ont en commun le fait que le prédicat nominal est en construction directe. Tandis que dans la construction présentée par la structure (5), le prédicat nominal est en construction prépositionnelle, tandis que le complément est en construction directe. Les formes à complément prépositionnel ont été, à leur tour, subdivisées en plusieurs classes.

PHRASE SIMPLE OU À COMPLÉMENT COMPLEXE ?

La question de savoir s'il s'agit de phrase simple ou de constructions complexes s'impose, presque de manière automatique, à chaque fois qu'il est question d'analyser une structure à verbe support de type N_0 V_{sup} N prép N_1. Ainsi :

1- Luc prend Max sous sa direction
2- Luc prend la décision de partir
3- Luc prend Max en croupe

On constate que lorsque le V-n se trouve en position de prédicat nominal, le cas de décision, est rattaché à un verbe, ici décider, le statut de chaque argument N_0 et N_1, est prédéfini par leur relation dans la construction verbale.

- Luc décide de partir

La question qui se pose est de savoir si les constructions avec le verbe support *prendre* lorsqu'elles peuvent être reliées soit à des phrases ayant d'autres supports : être, avoir, etc., présentent ou non entre elles des relations de dépendances par différents procédés syntaxiques. Nous avons vu, dans le cas du complément de N, qu'il était possible de relier les phrases entre elles par relativisation.

On voit que la réponse à cette question n'est pas aussi simple, et demande une attention particulière. D'autant plus que la diversité des compléments prépositionnels ne permet pas, au premier abord, de savoir s'il est question ou non de structure simple ou de

phrase complexe. Nous reviendrons sur cette question dans différentes parties de ce travail, et plus particulièrement lors de l'analyse des constructions prépositionnelles en *prendre*.

RELATION DE N À N_0 : FORMATION D'UN GN

Si l'on compare les phrases, on constate que le substantif *chapeau* peut avoir un complément de N_{hum}, mais non pas *décision* :

- Max prend (un chapeau + une décision)
- Max a pris (le chapeau de Chloé + *la décision de Luc)

Le substantif *chapeau* prend un complément de N_{hum} qui est différent du N_0, tandis que décision nécessite la coréférence de ce complément avec le sujet du verbe *prendre*.

- Max a pris sa décision

La possibilité ou non d'avoir un complément nominal est un critère pour savoir si le N_0 est à la fois le sujet du verbe support et du prédicat nominal ou non. Si l'on remplace *prendre* par un autre verbe comme *apprécier*, par exemple, il devient possible d'avoir un complément prépositionnel de N_{hum}.

- Luc apprécie (cette décision + la décision de Max)

Dans cette dernière phrase, Luc est le sujet du verbe apprécier, mais il ne peut pas être l'agent du prédicat nominal. D'ailleurs, il est possible de restructurer, dans ce cas, le verbe support *prendre*, par relativisation.

- Luc apprécie la décision qu'a prise Max

Parallèlement à cette construction, il est possible d'avoir Max en position de sujet du prédicat nominal à l'intérieur de la complétive enchâssée après le verbe apprécier.

- Luc apprécie que Luc prenne cette décision
- Luc apprécie le fait que Max ait pris rapidement sa décision

De la même façon, la relation du sujet entre N_0 et *prendre* peut être confirmée par la possibilité d'avoir la structure passive :

- Luc apprécie la décision prise par Max

où Max est en position de complément d'agent. On peut dire, donc, que le N_0 entretient une relation de sujet à la fois avec le support *prendre* et le N prédicatif. Cette relation est l'un des critères qui permettent de distinguer les N prédicatifs des noms non prédicatifs.

DISTINCTION ENTRE N PRÉDICATIFS ET N NON PRÉDICATIFS

Cette question s'impose, vu la diversité des constructions dans lesquelles entre le verbe *prendre*. L'étendue du lexique dicte un choix à faire. S'il n'y a pas la moindre

hésitation à considérer des substantifs comme *décision* et *courage* comme des N prédicatifs, cependant, il est difficile de savoir ce qui distinguerait *train* de *chapeau.*

- Luc a pris (une décision + du courage + le train + le chapeau)

Nous avons vu que l'opposition N concret / N abstrait n'est pas un paramètre fiable pour une classification du lexique. Afin de bien distinguer les N prédicatifs des N non prédicatifs, il est nécessaire d'appliquer à l'élément nominal un ensemble de critères syntaxiques.

- prends (le livre + un morceau de tarte)
- *prends (le large + une gifle + plaisir)

D'autant plus, dans la valeur lexicale pleine, le verbe peut être remplacé — lorsqu'il exprime un processus ou une action — par la proforme faire, alors que celle-ci est incompatible quand le verbe en question — devient support du N prédicatif — avec sa valeur délexicalisée.

- Luc (mange + prend) une pomme et Max en fait autant
- *Luc prend (une gifle) et Max fait autant

En outre, lorsqu'un verbe est employé comme support, il permet même d'actualiser des N concrets par métonymie et qui deviennent de ce fait des vrais N prédicatifs :

- Luc a pris un verre = Luc a bu un verre
- Luc a pris son repas = Luc a mangé son repas

Même s'il est parfois difficile de mettre en relation les propriétés syntaxiques qui permettent l'actualisation de relations d'interdépendances entre certaines catégories grammaticales et les entrées lexicales, le choix d'une métonymie plutôt qu'une autre dépend en fait exclusivement de considérations extralinguistiques.

LES CONTRAINTES SUR LES DÉTERMINANTS

La détermination semble être un des critères qui permettent de distinguer les N libres des N non libres. Le mot *chapeau* dans la phrase suivante n'impose pas de contrainte particulière à sa détermination.

- Luc a pris (un + ce + le + mon + son) chapeau

Même le déterminant possessif qui vient de la structure N de N n'impose pas de contrainte de coréférence entre le sujet de *prendre* et le complément du nom *chapeau.* Son, par exemple, peut avoir deux référents différents dans cet exemple, puisqu'il peut être coréférent à Luc, comme il peut renvoyer à une tierce personne, Marie, par exemple. Contrairement à cela, dans une phrase à support, le N prédicatif impose la coréférence entre le possessif et le sujet du verbe *prendre.*

- Luc a pris (sa + *ta + *ma) décision

Cette contrainte sur la coréférence s'impose même lorsque N prédicatif n'émane pas nécessairement de N_0.

- Luc prend (ses + *tes + *mes) ordres de Max

Même si les ordres sont de Max, pourtant l'adjectif possessif doit avoir le même indice que le sujet N_0. Dans ce cas, il n'est pas possible, par exemple, d'avoir le déterminant défini *les* :

- *Luc prend les ordres de Max

LA QUESTION PAR *QUE* PORTANT SUR LE N_1

Le test du questionnement par *que* semble être déterminant pour distinguer les prédicatifs des N non prédicatifs. Pour les verbes pleins, le test question-réponse par *que* permet de savoir si N_1 assume la fonction de complément d'objet. Ainsi, pour le verbe *prendre* on a :

- Luc a pris son chapeau
- Qu'a pris Luc ?
- son chapeau

Par contre, dans son emploi de verbe support, le test question-réponse produit un énoncé erroné.

- Luc prend du plaisir
- Que prend Luc ?
- du plaisir

Luc prend ses ordres de Max
- Que prend Luc de Max ?
- ses ordres

Luc prend ce cours sous sa responsabilité
- Que prend Luc sous sa responsabilité ?
- ce cours

Luc prend l'air sur la terrasse
- Que prend Luc ?
- l'air sur la terrasse

LES DÉTERMINANTS DE N : ARTICLE INDÉFINI OU PAS

La question de la détermination est primordiale pour savoir si, dans la construction à verbe support, le prédicat nominal est un N libre ou s'il est question d'une structure figée. D'ailleurs, la possibilité ou non de la variation de la détermination est en relation avec l'application ou non d'un certain nombre de transformations syntaxiques. Ainsi, si nous considérons les phrases suivantes :

- Luc prend (un + ce + ton + son) chapeau
- Luc prend (un + un beau + un certain) livre de la bibliothèque

Le déterminant, du complément d'objet *chapeau* et *livre*, n'est pas contraint syntaxiquement. Cependant, ceci n'est pas généralement le cas avec les N prédicatifs dans les constructions à support. Nous avons vu que les constructions en *prendre* imposent par contre des contraintes à la détermination du prédicat nominal.

- Luc a pris une (E + fausse) information de Max

On voit bien que le substantif, *information*, après le support *prendre* peut aussi avoir des déterminants variés. Cependant, la contrainte qui pèse sur ce type de prédicats est la nécessité de coréférence entre le sujet N_0 de *prendre* et le complément prépositionnel N_1. Ce dernier étant à son tour le sujet du N prédictif. Ceci est confirmé par le fait que le possessif avec le V-n doit renvoyer nécessairement à N_0.

- *Luc prend Max sous (ma + ta) responsabilité
- *Luc prend Max sous (mes + tes) ordres

Cette propriété de coréférence est particulière à l'emploi des verbes supports par opposition à l'emploi des verbes ordinaires.

- Luc a acheté (mon + ton + son) journal
- Luc a mangé (ma + ta + sa) soupe

Nous avons vu que la contrainte de coréférence était l'élément essentiel pour la formation de complément N_{hum} avec un N prédicatif par opposition à un N non prédicatif.

- *Luc a pris la décision de Max
- Luc a acheté le journal de Max
- Luc apprécie la décision de Max

Le choix des déterminants possibles pour un complément est un critère essentiel qui est en relation avec la possibilité ou non de l'application de certaines transformations. Ainsi, on peut mettre en évidence la relation qui existe entre le fait qu'un N prédicatif prenne le déterminant indéfini et la transformation relative. Cette relation est validée par les constructions à support *prendre* que nous analysons. Ce lien entre la détermination et la possibilité de la transformation relative est extensible à tout le lexique.

- Luc prend conseil auprès de Max
- le conseil que Luc prend auprès de Max

Nous avons vu, auparavant, que la relativisation permettait la formation de GN de forme : N de N_0

- Luc est surpris du dessus que Max prend sur Chloé
- Luc est surpris du dessus de Max sur Chloé

Cependant, la formation de la relative n'est pas systématique, puisque certains substantifs se prêtent mal à une telle transformation, malgré le fait que le déterminant n'est pas totalement figé.

- Luc prend (un + un grand) plaisir à lire ce livre
- Le plaisir que Max prend à lire ce livre [est particulier]
- *Son grand plaisir à lire ce livre [est particulier]

Le fait que le N prédicatif accepte ou non le déterminant indéfini peut être un critère décisif dans la classification du lexique. En effet, il permet de distinguer les N libres, qui n'ont pas généralement de contraintes particulières sur la variation de leurs déterminants, des N non libres, et qui devient de ce fait une construction figée. La question de la variation de détermination est un facteur qui autorise ou non la formation de groupes nominaux de forme le N de N_0.

- Luc prend (E + un particulier) plaisir
- le plaisir que prend Luc
- son plaisir

Le prédicat nominal, *plaisir*, n'accepte ni le déterminant indéfini, ni le déterminant défini.

- *Luc prend (un + le) plaisir

Cependant, l'utilisation de l'article indéfini nécessite la présence d'un modifieur. Le fait que plaisir prenne le déterminant Un - Modif est le signe qu'il est question d'un substantif libre. La preuve en est la possibilité d'utiliser d'autres déterminants.

- Luc prend (beaucoup de + du) plaisir

Dans la présentation des tables, les déterminants indéfinis — un + Un - Modif, du + des — sont notés dans les colonnes. Ce qui permet de prendre en compte des cas où le déterminant indéfini est obligatoirement accompagné d'un modifieur. Ceci permet, par exemple, de distinguer les différents emplois, du prédicat.

1a - Luc prend l'air
1b - Luc prend de l'air
1c - *Luc prend un air
2a - Luc prend un air sévère
2b - *Luc prend l'air sévère

Dans (1a) et (1b), le substantif *air* renvoie à l'oxygène. D'ailleurs, dans ce cas, il est possible d'utiliser un quantifieur.

1d - Luc prend une bouffée d'air

Parallèlement, il est possible d'avoir le déterminant Du - Modif.

1e - Luc prend de l'air pur

1f - Luc prend une bouffée d'air pur

Or, lorsqu'il est question de l'apparence, il n'est pas possible d'avoir Du - Modif.

2c - *Luc prend de l'air sévère

De même, la quantification n'est plus possible.

2e - *Luc prend une bouffée d'air sévère

Pour ce qui est des constructions figées, nous les avons classées dans des tables particulières. Elles se caractérisent par le fait que le déterminant est invariable, et que chaque entrée lexicale ne prend qu'un déterminant spécifique. Ce caractère fait du Dét une partie intégrante du prédicat nominal. Pour ce type de tables, nous n'avons consacré qu'une seule colonne à Dét, et que nous avons noté en toutes lettres. Nous avons ainsi :

- Luc prend (le large + son pied + terre)

Cependant, pour les constructions figées, nous n'avons pas spécifié pour chaque déterminant s'il était ou non accompagné d'un modifieur. Ainsi, dans les tables, nous avons regroupé les N simples, (comme large + pied ...) et des N composés à modifieur adjectival AN ou prépositionnel N prép N. Ce paramètre d'extension du prédicat nominal a été pris en considération dans la subdivision des constructions figées, comme d'ailleurs pour les autres tables, en fonction de la nature du substantif. Ainsi, nous avons pour la table N *prendre* C :

- le cas ou N est seul
- Le cas où N est accompagné nécessairement d'un adjectif
- Le cas où N est suivi obligatoirement d'un modifieur prépositionnel.

LES N LIBRES : LES CONSTRUCTIONS PRENDRE N SOURCES DE GN

Nous avons discuté auparavant des critères qui permettent de distinguer les N prédicatifs libres des N prédicatifs non libres. L'un des critères que nous avons retenu était la possibilité d'avoir ou non un déterminant indéfini, exception faite des constructions où il est figé.

- Luc a pris une gifle

Cette propriété fait que les prédicats nominaux à supports fonctionnent comme n'importe quel substantif dans n'importe quelle construction, puisqu'il peut être aussi bien en position sujet qu'en position complément.

On constate donc que le verbe *prendre* reçoit dans la structure de base ou libre une valeur de verbe plein, qui reste constante puisqu'elle dépend de son statut lexical. Cette valeur est déterminée par sa structure distributionnelle facilement définissable et qui correspond à sa structure la plus étendue ou chaque argument est utilisé pour ainsi dire dans son sens concret ou premier. Tandis que dans les emplois supports, le verbe, malgré

la parenté de la structure dans laquelle il s'insère, n'a presque plus rien gardé du sens premier, et qu'il s'est vidé presque de son sens, et acquiert des valeurs sémantiques différentes dépendantes des N prédicatifs auxquels il s'associe et dont il est le support. Dans ce processus d'extension d'utilisation, le verbe plein devient presque transparent à la relation prédicative, à l'image des opérateurs aspectuels. Cependant, il garde, en partie, souvenir de sa combinatoire libre.

Ainsi, dans ce processus de désémantisation et de délexicalisation, le verbe support même s'il voit modifier les contraintes distributionnelles qu'il imposait ou qui le reliaient à ses arguments dans les constructions libres, il ne rompt pas définitivement pour autant sur le plan syntagmatique et sémantique avec son passé historique et garde une certaine image de ses emplois.

Le plus souvent, et il n'y a pas très longtemps, on associait, de manière presque systématique, les verbes supports aux constructions contenant un nom d'action dérivé d'un verbe comme dans l'exemple suivant :

- Luc a décidé de partir
- Luc a pris la décision de partir

Ainsi, *prendre* est considéré comme un support de nominalisation, comme on parle dans d'autres cas de supports d'adjectivation. Cependant, on s'est rendu compte de l'existence de structures syntaxiques dans lesquelles le verbe assume le support de prédication d'un nom, sans que ce dernier puisse être relié de quelques manières, sauf par des relations ad hoc, à une base verbale ou adjectivale. D'autant plus que ces substantifs ne sont pas nécessairement des N abstraits. Ainsi dans :

- Luc a pris (le train + l'avion + le métro)

De même, l'existence de structures dont au moins un élément est figé :

- Luc a pris (le large + la mer)

Nous constatons alors l'existence d'un certain continuum entre les emplois libres, les emplois supports et les emplois figés. Ce continuum qui au départ n'était pas prévisible, mais grâce à l'étude systématique du lexique et à l'établissement de classe de noms, il est possible de justifier et d'expliquer les relations qui existent entre les différents emplois.

Ainsi, donc, grâce à ce continuum, on peut rendre compte à la fois de la diversité des emplois et des constructions syntaxiques, des conditions dans lesquelles ils acquièrent des sens diversifiés, puisque l'on peut dire qu'il existe autant d'entrées lexicales pour le même lexème. Cette diversité résulte à la fois du fait qu'il garde partiellement des traces sémantiques de son emploi libre, et en perd lorsqu'il contribue à la construction du sens des éléments prédicatifs auxquels il s'associe.

Ainsi, dans les emplois supports, l'élément en question en se délexicalisant, se neutralise pour n'être qu'un élément neutre de la prédication. Cette neutralité s'accentue dans le cas des constructions figées puisque l'élément en question perd totalement toute

aliénation avec son emploi libre comme dans le cas de l'expression *prendre* le taureau par les cornes.

Ce processus de neutralisation ne peut être déterminé de manière claire que dans le cadre du lexique-grammaire en analysant toutes les classes de noms, et les supports qui peuvent leur être associés. C'est au terme de ce travail qu'on pourrait rendre compte de la neutralisation des valeurs sémantiques de chaque mot, des phénomènes de la synonymie, des mécanismes de métaphorisation et de métonymie. Il serait intéressant de se pencher sur les phénomènes syntaxiques et sémantiques

La particularité de cette classe de mots — limitée en nombre et dont l'utilisation et la fréquence d'emploi sur le plan statistique est très importante — qui peuvent devenir des éléments neutralisés dans des emplois supports est de subir des changements sémantiques. Chaque mot entre son emploi premier, qui correspond à la définition sémantique du dictionnaire, l'emploi support et l'emploi métaphorique déploie toutes ses différentes significations comme s'il manifestait dans chaque construction l'une de ses différentes facettes. L'actualisation dans le discours d'un mot relève à la fois du même et du différent. Du même, puisque l'entité garde une trace pour ne pas dire un sème de son sens primaire comme s'il gardait en mémoire la raison de son existence ou la signification pour laquelle il a été créé et qu'il doit assumer. Du différent, puisque le processus de son actualisation dans un énoncé l'amène à des contraintes combinatoires, qui lui font perdre petit à petit et partiellement son sens jusqu'à le neutraliser complètement. À chaque fois qu'il se combine et s'associe à un N prédicatif, on a le sentiment qu'il perd ses propriétés syntaxiques et sémantiques puisqu'il constitue avec lui à chaque fois une entité nouvelle, ce qui a fait que plusieurs auteurs traitent les constructions supports comme des locutions ou expressions figées ou figurées. Certains parlent d'une coalescence verbo-nominale, puisque aussi bien le verbe que le nom prédicatif perdent chacun une partie de ses propriétés pour faire place tous les deux à une signification nouvelle qui n'est pas nécessairement la somme des significations de l'un et de l'autre. Chaque association, chaque actualisation renverse la ou les significations des items, ce qui rend difficile toute classification du lexique, vu le nombre des possibilités combinatoires et plus particulièrement des noms prédicatifs.

Les supports ont donc cette faculté de mutation et de transformation sémantique et syntaxique puisque leur entité mue et change systématiquement à chaque combinaison. Ils acquièrent des significations différentes à la fois par association à d'autres éléments lexicaux et par mimétisme et analogie puisqu'ils intègrent partiellement des significations de l'élément auquel ils sont combinés. Par conséquent, ils se mettent à ressembler à d'autres entités — généralement de la même catégorie et ayant la même fonction et parfois même différente, ce qui peut être relié au phénomène de synonymie — de sorte qu'ils peuvent assumer et prendre leur place dans un processus de substitution lexicale et actualiser au moins une des significations de ce dernier.

Les supports peuvent avoir entre autres comme fonction d'établir des relations facilement définissables et repérables avec les N prédicatifs qu'ils actualisent dans le discours. Ces relations peuvent se déduire de l'établissement de classes d'objets (Cf. G. Gross), ce qui permet de justifier et d'expliquer comment ils parviennent à actualiser l'une des significations de l'élément supporté, mais même si les résultats sont probants, il en reste que la distribution et le fonctionnement à l'intérieur de la même classe n'obéit pas de

manière systématique et mécanique aux mêmes règles combinatoires vu certaines comptabilités.

Comme on peut le constater, le figement de ces constructions ou de la séquence (verbe support - N prédicatif) incombe au fait des contraintes de détermination. Le type de déterminant, la possibilité ou non de le varier font qu'il devient un élément central de la construction puisque le figement dépend essentiellement de lui. C'est dans ce sens qu'on préconise la séparation des constructions figées des constructions à supports et des constructions libres. N'empêche qu'il y a un certain continuum entre les différentes constructions puisque «les règles que subissent les expressions figées sont exactement les règles de la syntaxe des phrases libres et ce aussi bien pour leurs parties libres que pour leurs parties figées» (M. Gross, 1988).

RÉFÉRENCES

BOONS, J.-P., GUILLET, A. et Ch. LECLERE (1976) : *La structure des phrases simples en français : construction intransitive*, Genève, Droz, 377 p.

GIRY-SCHNEIDER, J. (1978) : *Les nominalisations en français : l'opérateur faire dans le lexique*, Genève, Droz, 353 p.

GIRY-SCHNEIDER, J. (1978) : *Les prédicats nominaux en français : les phrases simples à verbes supports*, Genève, Droz, 396 p.

GROSS, G. (1989) : *Les constructions converses du français*, Genève, Droz.

GROSS, G. (1994) : «Classe d'objets et traitement de la synonymie», Ibrahim, A. H. (dir.), *Supports, opérateurs, durée*, Annales de l'Université de Besançon 516, Série Linguistique et Sémiotique, vol. 23, Paris, Les Belles Lettres, 268 p.

GROSS, G. (1994b) : «Classe d'objets et descriptions des verbes», *Langages*, 115, Paris, Larousse.

GROSS, M. (1975) : *Méthodes en syntaxe*, Paris, Hermann.

GROSS, M. (1976) : «Sur quelques groupes nominaux complexes», *Méthodes en grammaire française*, Paris, Klincksiek.

GROSS, M. (1981) : «Les bases empiriques de la notion de prédicat sémantique», *Langages*, 63, Paris, Larousse.

GROSS, M. (1982) : «Une classification des phrases figées du français», *Revue Québécoise de Linguistique*, XI-2, Montréal, Presses de l'Université de Québec.

GROSS, M. (1986) : «Les nominalisations d'expressions figées», *Langue française*, 69, Paris, Larousse.

GROSS, M. (1988) : «Les limites de la phrases figée», *Langages*, 90, Paris, Larousse.

HARRIS, Z. S. (1970) : *Papers in Structural and Transformational Linguistics*, Dordrecht, D. Reidel.

IBRAHIM, A. H. (1984) : «Sur le statut de quelques accidents syntactico-sémantiques», De la syntaxe à la pragmatique, vol. 8 de *Linguisticae Investigationes Supptementa*, Amsterdam/Philadelphia, John Benjamins.

IBRAHIM, A. H. (1993) : «La déviance de la suffixation en français est-elle structurelle ?» *TRANEL,* Traitement des données linguistiques non standard, 20, Neuchâtel, Université de Neuchâtel.

IBRAHIM, A. H. (dir) (1994) : *Supports, Opérateurs, Durées*, Annales de l'Université de Besançon 516, Série Linguistique et Sémiotique vol. 23, Paris, Les Belles Lettres, 268p.

LE PROJET NADIA-DEC : VERS UN DICTIONNAIRE EXPLICATIF ET COMBINATOIRE INFORMATISÉ ?

Gilles SÉRASSET

GETA-CLIPS-IMAG (UJF & CNRS), Grenoble, France

INTRODUCTION

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème des ressources lexicales et linguistiques s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en termes de quantité, de qualité et de complexité. La complexité et la diversité des informations requises augmentent avec les exigences des outils de TALN ainsi qu'avec le développement de nouvelles applications (humaines ou machinales). Si la récupération (semi)automatique d'information lexicale est une piste, elle ne pourra remplacer la création manuelle de dictionnaires.

Nous nous sommes donc intéressé à la construction d'outils pour lexicographes et lexicologues. Afin d'avoir une bonne compréhension des problèmes qui se posent, nous avons décidé d'informatiser un dictionnaire complexe, contenant de nombreuses informations structurées, le *Dictionnaire explicatif et combinatoire du français contemporain* (DEC). Le DEC étant un travail de lexicologie, il ne s'agit donc pas à proprement parler d'un dictionnaire, mais plutôt d'un ensemble d'entrées destinées à illustrer une théorie linguistique. Ce ne sont donc pas les données que l'on a informatisées, mais le processus de rédaction de ces données.

Les travaux menés au cours du projet NADIA-DEC s'appuient d'une part sur le système SUBLIM (Sérasset 1994) défini au laboratoire GETA-CLIPS de l'Université Joseph Fourier (Grenoble I) et, d'autre part, sur les travaux de lexicologie menés par l'équipe d'Igor Mel'cuk (Mel'cuk et al. 1995) au laboratoire GRESLET de l'Université de Montréal. Il a reçu le soutient du réseau LTT de l'AUPELF-UREF et des ministères français et canadiens des affaires étrangères.

Ce projet pose des problèmes informatiques et linguistiques sérieux. L'aspect évolutif de la structure du DEC impose la nécessité de fournir des outils informatiques adaptables. La nécessité de formalisation des informations peut conduire à différentes stratégies de représentation des informations.

Nous montrons les différentes étapes de l'informatisation du DEC en donnant tout d'abord la structure interne des informations lexicales. Nous donnons ensuite un aperçu des outils et méthodes utilisés pour la création et la validation d'un DEC informatisé.

LE PROJET NADIA-DEC

Objectifs

Depuis 1994, le GRESLET (Université de Montréal) et le GETA-CLIPS (Université Joseph Fourier - Grenoble I) travaillent ensemble sur le projet NADIA-DEC, soutenu par le réseau LTT de l'AUPELF-UREF.

L'objectif de ce projet est l'informatisation du *Dictionnaire explicatif et combinatoire du français contemporain* (DEC) créé par Igor Mel'cuk. Cette informatisation se base sur les travaux préalablement effectués au GETA (Sérasset 1994), et répond aux contraintes suivantes :

- **Fidélité linguistique** : les structures informatiques utilisées doivent rester proches des structures linguistiques que l'on souhaite représenter;

- **Généricité** : les outils construits doivent pouvoir être utilisés pour d'autres structures;

- **Adaptabilité** : les outils informatiques doivent pouvoir évoluer en même temps que la structure informatique.

Le DEC étant en constante évolution, il nous est très vite apparu important d'informatiser non seulement les données existantes, mais surtout sa production. Nous avons donc développé des outils d'éditions spécialisés pour le DEC ainsi que des outils de récupération des informations déjà décrites disponibles sous forme de fichiers Word™.

L'approche utilisée ne remet pas en cause la structure linguistique que l'on peut trouver dans le DEC. La structure informatique du DEC doit permettre, au minimum, de re-générer à l'identique les fichiers Word™ utilisés pour la version papier. Aussi, toutes les informations sont présentes, et ce même si elles ne sont pas structurées. Il est ainsi toujours possible, au fur et à mesure que l'on avance dans ce projet, d'augmenter la structuration des données sans avoir à reprendre l'ensemble du processus de récupération à partir des fichier Word™.

À cet égard, le projet NADIA-DEC se distingue des autres projets d'informatisation du DEC, qui se basent a priori sur une structure informatique simplifiée et qui n'informatise que le sous-ensemble de données commun entre le DEC et cette structure.

Enfin, les données du DEC ne sont pas récupérées dans le but d'une utilisation informatique particulière. Nous estimons que cette indépendance par rapport à l'usage qui sera fait des données nous permet de garantir la complétude des informations récupérées.

Méthodologie

Nous avons distingué plusieurs tâches pour accomplir le projet NADIA-DEC :

- définition d'une structure informatique pour le DEC,
- récupération des informations existantes sous cette forme structurée,
- construction d'un éditeur spécialisé pour cette structure (l'éditeur DECID),
- exportation des données structurées vers différentes formes.

Ainsi, notre méthodologie peut être résumée par le schéma suivant :

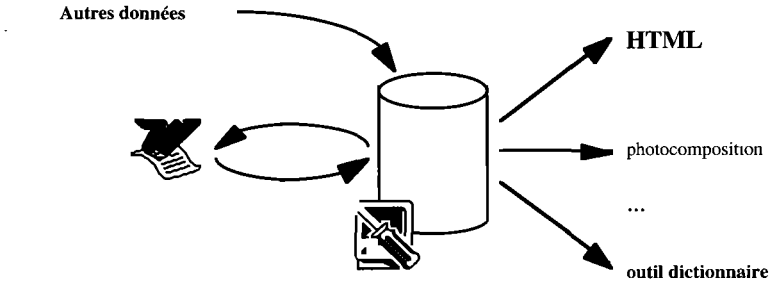


Figure 1 : Méthodologie de création d'un DEC informatisé

L'ÉDITEUR DECID

DECID est un éditeur spécialisé pour l'édition du DEC. Il offre de nombreuses fonctionnalités pour aider le lexicographe. Sa conception et son implantation ont été effectuées dans un souci de simplicité et de convivialité.

En utilisant l'éditeur DECID, le lexicographe crée ou modifie, en direct, une structure informatique. Pourtant, l'interface a été conçue pour lui donner l'impression de travailler, comme auparavant, sur le DEC tel qu'il est publié. Aussi, la visualisation des données est-elle très proche de celle qui est utilisée dans la version papier.

Le lexicographe dispose d'une fenêtre principale lui donnant la liste des vocables et des lexies du fichier en cours d'édition (figure 2). Le second type de fenêtre présente et permet d'éditer une lexie.

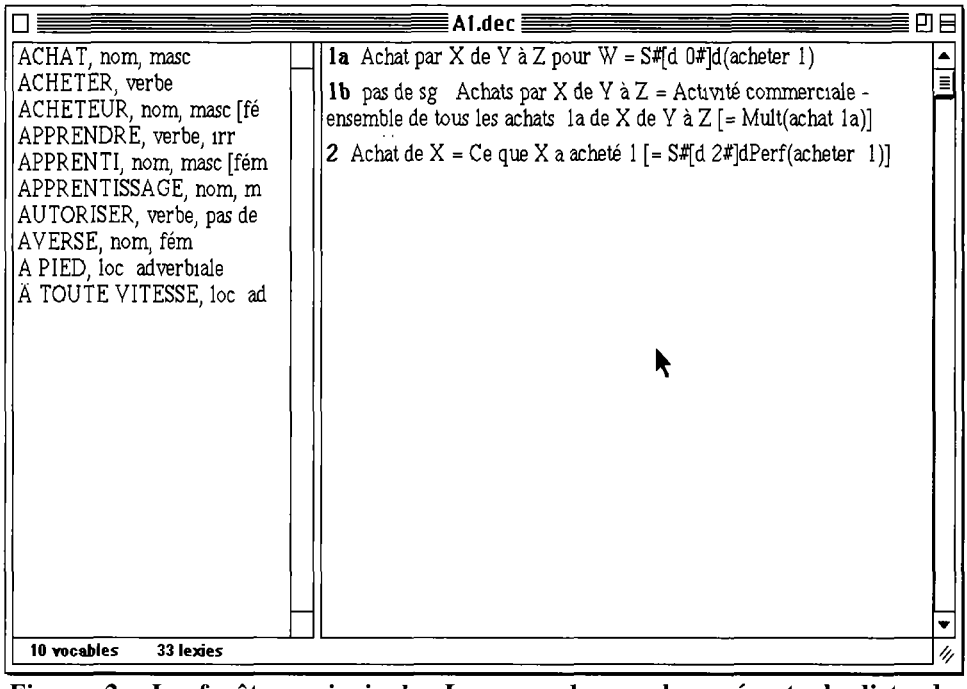


Figure 2 : La fenêtre principale. La zone de gauche présente la liste des vocables du fichier en cours d'édition. La zone de droite présente la liste des lexies du ou des vocables sélectionnés

La fenêtre de lexie permet d'éditer le vocable, le numéro de la lexie, les informations morphologiques, la définition et les exemples de manière très simple. La zone de régime n'est pas encore traitée par l'éditeur DECID.

Les fonctions lexicales apparaissent sous une forme très proche de la forme papier. Mais leur édition a été rendue très aisée par l'éditeur DECID. En effet, auparavant, le lexicographe devait, pour éditer une fonction lexicale, utiliser des indices, des exposants, des changements de fontes. Il devait faire attention à la correction du nom de la fonction, bien mettre le premier caractère en majuscule et le reste en minuscule... Tous ces soucis ont maintenant disparu lorsqu'on utilise l'éditeur DECID. En effet, le lexicographe se contentera de taper : `perm1incepreal3+'usual` pour voir se dessiner la fonction : $\mathbf{Perm}_1\mathbf{IncepReal}_{3\subset}^{\text{usual}}$.

Le lexicographe pourra ensuite sauver son travail sous forme structurée ou alors sous forme d'un fichier RTF (Rich Text Format) qu'il pourra ensuite utiliser directement en Word.

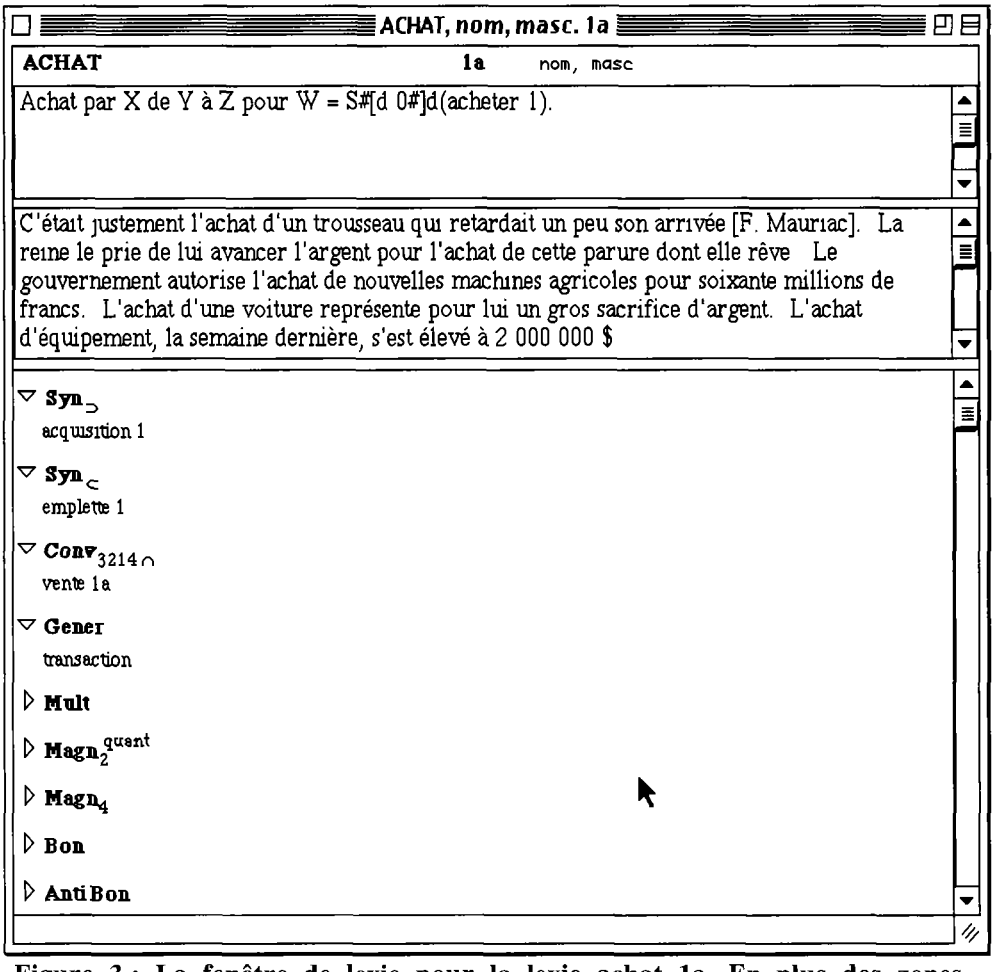


Figure 3 : La fenêtre de lexie pour la lexie achat 1a. En plus des zones de définition et d'exemple, on remarque la zone des fonctions lexicales

RÉCUPÉRATION DES DONNÉES EXISTANTES

En plus de l'éditeur DECID, nous avons développé un outil de récupération des données publiées du DEC. Cet outil part d'un fichier en RTF (Rich Text Format) généré à partir des fichiers Word™ qui ont été utilisés à l'origine pour la création du DEC papier.

Cette récupération n'a pu se faire que semi-automatiquement. Les fichiers en cours de récupération devant être corrigés pour être récupérables. Fort heureusement, les fichiers avaient, dès le départ été créés en utilisant des styles cohérents pour les différents paragraphes décrivant une entrée (définition, régime, etc.). Sans cela, la récupération n'aurait pu avoir lieu.

Certaines difficultés sont dues à l'outil Word™ utilisé. L'absence totale de documentation du format RTF nous a obligé à produire des outils ad hoc. De plus, pour des raisons encore assez obscures, des fichiers d'apparence identiques ont des descriptions RTF différentes. Ainsi, deux paragraphes successifs ayant le même style peuvent apparaître soit comme deux paragraphes indépendants (l'information de style est donnée pour chaque paragraphe) ou comme deux paragraphes identiques (la définition de style n'est donnée qu'au début du premier).

D'autres difficultés sont dues au DEC lui même. Le DEC a été conçu au départ sous une forme papier utilisable par un homme. Aussi, le DEC a été conçu avant tout par sa **présentation**. Aussi, une très grande importance a été accordée à la forme plus qu'à la structure. Ainsi, certaines erreurs dans les documents Word n'étaient pas détectées car elles n'étaient pas visibles sur papier. Par exemple, la définition et la liste d'exemples ont une même forme, mais sont représentées par deux styles différents. Néanmoins, on trouve souvent des erreurs dues à l'identité de forme de ces deux types d'information, l'humain pouvant très facilement faire la différence par le contexte.

Dans les tableaux de régimes, la présentation était contrôlée entièrement par le lexicographe. Ainsi, certaines lignes pouvaient être séparées par une marque de fin de paragraphe, un saut à la ligne, ou même par une succession de tabulations. Certaines valeurs pouvaient tenir sur plusieurs lignes. Dans ce cas, les valeurs apparaissent, dans le fichier séquentiel RTF, de manière entrelacée.

Ces erreurs de présentation et d'édition ont été réglées. Elles ont été fort heureusement assez mineures.

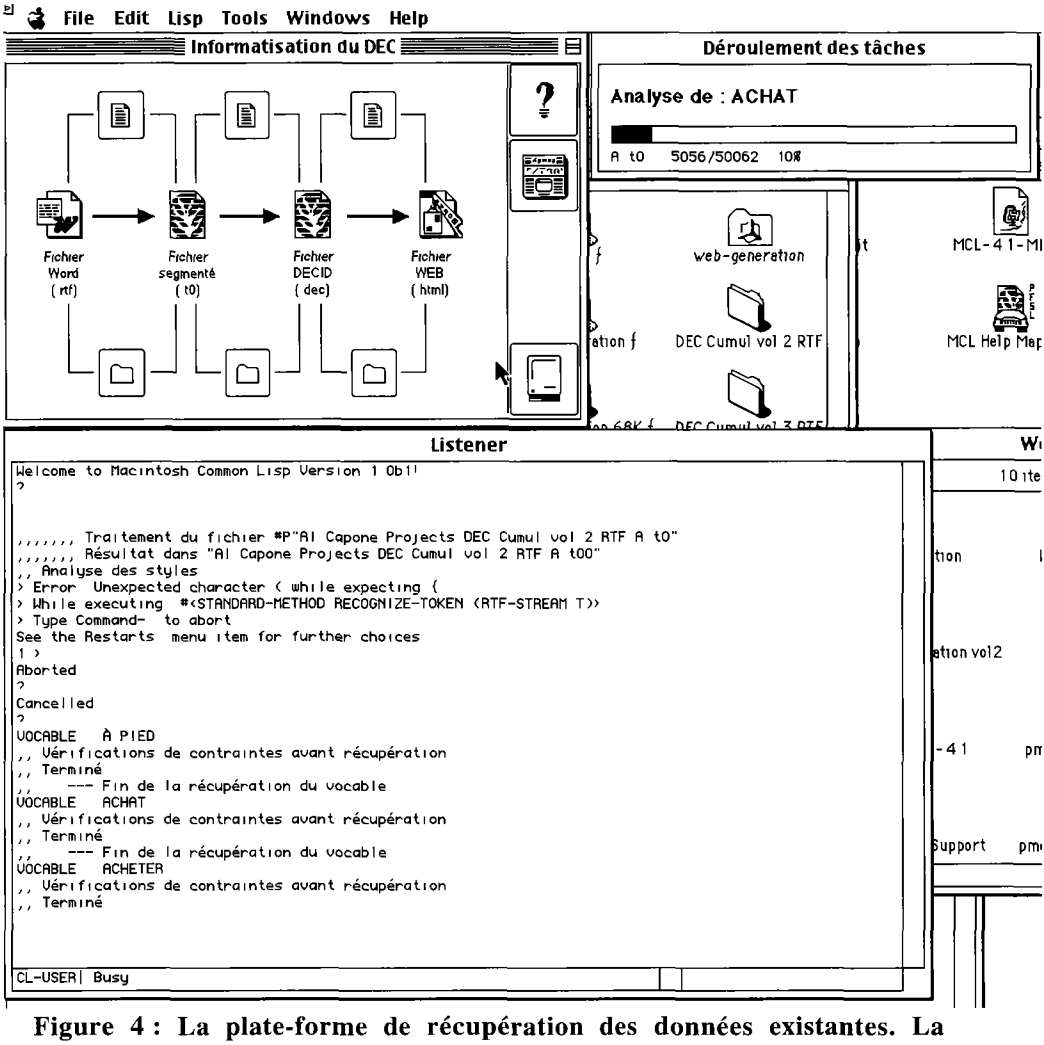


Figure 4 : La plate-forme de récupération des données existantes. La fenêtre en haut à gauche permet au récupérateur de déclencher les traitements. La fenêtre du bas donne des informations sur le traitement en cours

En utilisant la plate-forme de récupération (figure 4), le lexicographe déclenche la récupération d'un fichier RTF ou de tous les fichiers RTF présents dans un dossier. Ces fichiers sont analysés puis, pour chaque vocable détecté, un fichier est créé qui contient les informations sous forme structuré. Pour chaque vocable, le lexicographe peut demander un diagnostic qui lui sera utile pour savoir si les données ont été récupérées de manières satisfaisantes ou si une erreur s'est glissée dans la récupération sans que le processus ne se soit interrompu (figure 5).

L'outil de récupération aide le correcteur en donnant un diagnostic du vocable récupéré. Ainsi, par comparaison avec la version papier, il est aisé de voir ce qui n'a pas fonctionné et pourquoi.

```
;; Vocable      : ACHAT
;; Catégorie    : nom, masc.
;; Possède un tableau résumé : 3 résumés.
;; Pas de note.
;; Lexie    : 1a.
;;  Pas de connotations.
;;  Possède des informations de régime :
;;    Tableau à 4 colonne(s)
;;1 = X  | 2 = Y  | 3 = Z  | 4 = W
;;    3  |  1     |  2     |  2
;;    3 restriction(s) numérotées.
;;    6 exemples de réalisations
;;  Possède des fonctions lexicales :
;;    • 11 fonction(s) lexicale(s).
;;  Possède des exemples.
;; Lexie     : 1b.
;;  Pas de connotations.
;;  Possède des informations de régime :
;;    Tableau à 3 colonne(s)
;;    1 = X    |  2 = Y    |  3 = Z
;;       3     |     2     |     1
;;    2 restriction(s) numérotées.
;;    2 exemples de réalisations
;;  Possède des fonctions lexicales :
;;    • 14 fonction(s) lexicale(s).
;;  Possède des exemples.
;; Lexie     : 2.
;;  Pas de connotations.
;;  Possède des informations de régime :
;;    Tableau à 1 colonne(s)
;;    2 = X
;;       2
;;    0 restriction(s) numérotées.
;;    1 exemples de réalisations
;;  Possède des fonctions lexicales :
;;    • 3 fonction(s) lexicale(s).
;;  Possède des exemples.
```

Figure 5 : Le diagnostic de récupération d'un vocable

EXPLOITATION DES DONNÉES INFORMATISÉES

Les données ainsi informatisées ont été exportées sous forme HTML. Nous avons ainsi pu produire automatiquement un site Web complet présentant le DEC sous une **présentation** analogue à celle utilisée dans la version papier. Les figures 6 et 7 représentent une page du DEC vue par un navigueur standard.

Cette version HTML du DEC est, comme le DEC au format Word, destinée à un usage humain. Nous avons adopté une forme aussi proche que possible de la forme originale. Mais cette présentation pose différents problèmes. En effet, le format HTML ne permet pas, de manière simple, de préciser effectivement une forme. C'est le navigueur qui, en dernier ressort, effectue la présentation. Cela pose différents problèmes :

- les tables, n'apparaissent pas de la même manière suivant les navigueurs utilisés. Cela peut mener à des colonnes trop larges ou trop étroites;
- tous les navigueurs ne savent pas forcément interpréter et présenter les informations en indice ou en exposant.

Enfin, indépendamment de la compatibilité des différents navigueurs, la version HTML du DEC pose des problèmes intrinsèques. Ainsi, certains caractères sont propres au DEC (ex : k et l qui délimitent les locutions ou ' et " qui délimitent les sémantèmes). Ces caractères ne sont présents dans aucune fonte standard. Actuellement, HTML ne permet pas d'inclure une fonte ou une description de caractère qui soit portable. On peut

indiquer au navigueur d'utiliser une fonte particulière, mais celle-ci doit être présente dans le système du client. L'utilisation d'une image pour ces caractères peut être envisagée, mais le client n'aura pas une bonne présentation s'il décide de changer la taille des caractères affichés (l'image ne grandira pas en fonction).

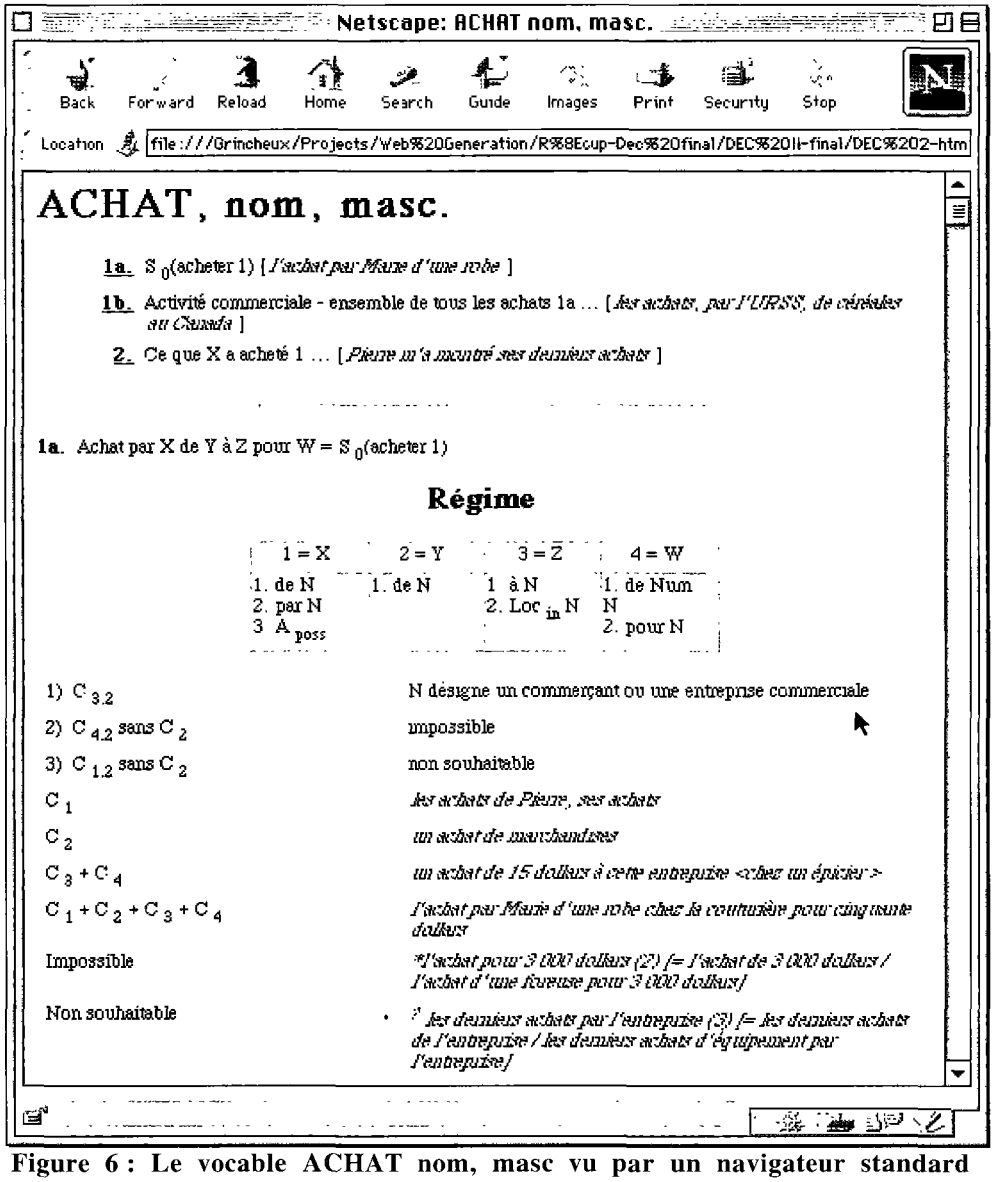


Figure 6 : Le vocable ACHAT nom, masc vu par un navigateur standard

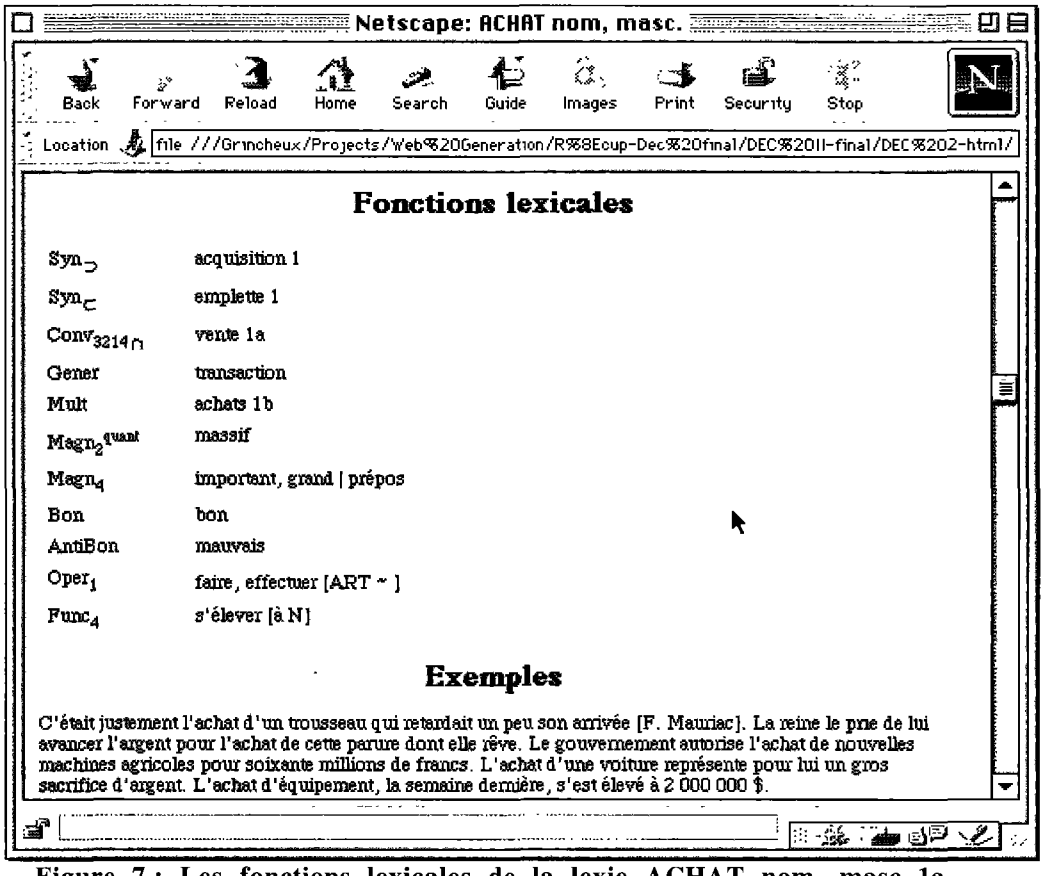


Figure 7 : Les fonctions lexicales de la lexie ACHAT nom, masc 1a

CONCLUSION

À l'occasion de l'action de recherche partagée NADIA-DEC, nous avons donc pu informatiser le *Dictionnaire explicatif et combinatoire du français contemporain*. Pour cela, nous avons récupéré semi automatiquement la totalité des entrées des volumes II et III du DEC.

Nous avons aussi créé un outil d'édition spécialisé pour le DEC. Cet outil offre des avantages certains aux lexicographes, mais il contraint trop la structure du dictionnaire et ne pourra être utilisé tant que l'édition du DEC se fera dans l'optique d'une recherche en lexicologie (structure en cours de définition).

Néanmoins, les travaux effectués seront très utiles pour un passage en phase de production du DEC ou d'un dictionnaire dérivé.

Ce travail a constitué une étape importante dans nos recherches sur des outils pour lexicographes. Elle nous a permis de mettre en œuvre nos méthodes sur un dictionnaire très complexe. Nous avons ainsi pu valider certains choix. Néanmoins, nous avons pu

voir que la construction d'outils spécialisés fige la structure informatique utilisée. Or, dès que l'on travaille avec des dictionnaires assez compliqués, la possibilité de remise en cause des structures informatiques en cours d'édition d'un dictionnaire est nécessaire.

Aussi, nous souhaitons orienter nos recherches sur des méthodes génériques de création d'outils spécialisés pour lexicographes. Cette généricité nous permettra d'offrir des outils évolutifs et rendra plus facile les recherches de lexicologie.

RÉFÉRENCES

MEL'CUK, Igor, CLAS, André et Alain POLGUÈRE (1995) : *Introduction à la lexicologie explicative et combinatoire*, coll. «Universités francophones» et «champs linguistiques», Louvain-la-Neuve, AUPELF-UREF et Duculot.

SÉRASSET, Gilles (1994) : *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*, Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, 194 p.

SÉRASSET, Gilles (1995) : «Informatisation du *Dictionnaire explicatif et combinatoire* : le projet NADIA-DEC», *Lexicomatique et dictionnairiques*, Actes des IV^es^ Journées scientifiques du réseau LTT, Lyon, 28-30 septembre 1995, pp. 205-215.

SÉRASSET, Gilles (1996) : «Un éditeur pour le Dictionnaire explicatif et combinatoire du français contemporain», *Journées lexique du PRC-CHM*, Grenoble, 13-14 novembre 1996, pp. 131-138.

SÉRASSET, Gilles (1997) : «Informatisation du Dictionnaire explicatif et combinatoire», *TALN'97*, Grenoble, 12-13 juin 1997, pp. 194-198.

SÉRASSET, Gilles et Étienne BLANC (1993) : «Une approche par acceptions pour les bases lexicales multilingues», *TA-TAO : recherches de pointe et applications immédiates*, Actes des III^es^ Journées scientifiques du réseau LTT, Montréal, 30 septembre-2 octobre 1993, pp. 65-84.

SÉRASSET, Gilles et Alain POLGUÈRE (1997) : «Outils pour lexicographes : application à la lexicographie explicative et combinatoire», *RIAO'97*, Montréal, 25-27 juin 1997, pp. 701-708.

TAO ET THÉORIES LINGUISTIQUES : INSTITUTIONS GRAMMATICALES*

Philippe BLACHE et Jean-Yves MORIN

LPL - CNRS, Aix-en-Provence, France et Université de Montréal, Canada

INTRODUCTION

Si l'on présuppose

(a) qu'il est inévitable en TALN (et plus spécifiquement en TAO) de tenir compte des propriétés fondamentales des langues naturelles (paradoxalement, ce ne semble pas être un truisme pour l'ensemble des chercheurs dans le domaine),

si, de plus, l'on admet que

(b) les théories et les descriptions linguistiques actuelles arrivent à identifier certaines de ces propriétés,

si, enfin, l'on croit que

(c) ces différentes théories et descriptions ne sont pas nécessairement contradictoires, mais présentent différents *points de vue* sur la realité,

on peut vouloir intégrer ces différentes descriptions, les rendre compatibles l'une à l'autre et les exploiter dans un système efficace et transparent.

Ce que nous nous proposons de construire, c'est un cadre général basé sur les notions de *dimensions*, d'*objets* et de *contraintes* permettant de réaliser à la fois l'œcuménisme théorique nécessaire et l'intégration des différents niveaux de représentation. On tentera donc de développer ce que nous appellerons des INSTITUTIONS GRAMMATICALES, qui fourniront de tels cadres généraux permettant de caractériser et de comparer la structure logique et calculatoire de ces théories et descriptions linguistiques.

* Une version préliminaire de ce travail a été présentée aux V^es^ Journées scientifiques du réseau LTT *La mémoire des mots* à Tunis, en septembre 1997. Ce travail a bénéficié de l'aide du réseau LTT à l'équipe «TAO et théories linguistiques» (Anne Abeillé, Philippe Blache, Jean-Yves Morin et Éric Wehrli) ainsi que d'une subvention de recherche du fonds FCAR du Gouvernement du Québec.

Il y a une quinzaine d'années, Goguen et Burstall (1984) introduisaient en logique informatique la notion d'*institution*.

> «This paper shows how some parts of computer science can be done in any suitable logical system, by introducting the notion of an **institution** as a precise generalization of the informal notion of a "logical system". A first main result shows that if an institution is such that interface declarations expressed in it can be glued together, then **theories** (which are just sets of sentences) in that institution can also be glued together. A second main result gives conditions under which a theorem prover for one institution can be validly used on theories from another; this uses the notion of an institution morphism. A third main result shows that institutions admitting free models can be extended to institutions whose theories may include, in addition to the original sentences, various kinds of constraints upon interpretations; such constraints are useful for defining abstract data types, and include so-called "data", "hierarchy", and "generating" constraints. Further results show how to define insitutions (sic) that mix sentences from one institution with constraints from another, and even mix sentences and (various kinds of) constraints from several different institutions.»
>
> [...]
>
> «Informally, an institution consists of
>
> - a collection of signatures (which are vocabularies for use in constructing sentences in a logical system) and signature morphisms, together with for each signature Σ,
> - a set of Σ-sentences,
> - a set of Σ-models, and
> - a Σ-satisfaction relation, of Σ-sentences by Σ-models
>
> such that when you change signatures (with a signature morphism), the satisfaction relation between sentences and models changes consistently.»
>
> (Goguen & Burstall 1984 : 221-222, l'emphase est dans l'original.)

En termes très simples, cela équivaut à dire que si l'on sait de quoi deux institutions A et B parlent (leurs signatures[1]), et que l'on sait traduire d'une institution à l'autre, on peut

---

[1] Rappelons qu'une signature est un couple Sig = <S, Σ> où

S est un ensemble (généralement structuré) de *sortes* ou *types* et

Σ est une famille de fonctions indexée par $S^* \times S$.

Pour σ $\sigma \in \Sigma_{as}$, on note

$\sigma : a \rightarrow s$

a est l'*arité* de σ (le nombre d'arguments de σ et leur type respectif)

s est la *sorte*

la paire <a, s> est le *rang* associé à σ.

En termes informels, une signature distribue les termes d'une théorie sur des ensembles ordonnés d'arguments typés (dans cette formulation, tous les termes sont des fonctions, les constantes sont des fonctions sans arguments, les prédicats à n arguments sont des fonctions de S^n dans V (un ensemble de valeurs de vérité). En fait, l'objectif de ce type de définition formelle (pour des lexiques abstraits) est tout à fait analogue à ce que font les définitions de type DEC (pour des lexiques concrets) qui attribuent à chaque lexie un type et une liste ordonnée d'arguments typés par des étiquettes sémantiques (cf. Mel'cuk et al. 1995; Miliçeviç 1997). Par exemple, pour VENDRE[_1] dans la syntaxe

combiner de façon cohérente les énoncés de A et ceux de B et déterminer des *conditions de collage* entre ces énoncés.

Pratiquement, il s'agit d'identifier les *prédicats théoriques* utilisés dans ces théories/descriptions, d'étudier leur sémantique descriptive (i.e., les DOMAINES d'entités qu'ils définissent/décrivent), leurs propriétés calculatoires (i.e., COMPLEXITÉ théorique et effective) et leur sémantique opératoire (représentabilité, directe ou indirecte, dans des implantations effectives) et d'élaborer des moyens d'intégrer les représentations correspondantes dans des environnements effectivement utilisables.

QUELQUES PROBLÈMES

Le volume et la disparité des connaissances linguistiques nécessaires en TAO pose de délicats problèmes d'ÉLABORATION et d'INTÉGRATION. En ce qui concerne l'ÉLABORATION, la plupart des travaux linguistiques menés depuis les trente dernières années portent plutôt sur l'interprétation théorique de l'analyse de phénomènes particuliers dans diverses langues que sur la construction de descriptions approfondies et couvrantes pour une langue en particulier[2]. Ceci tient d'une part à la nature extrêmement complexe des phénomènes linguistiques, qui ne se prêtent pas à une analyse globale rapide, ou même incrémentale, sauf dans certains domaines spécifiques (phonologie, morphologie, une partie du lexique) et d'autre part à des facteurs socio-historiques de développement de la discipline qui favorisent les travaux théoriques plutôt que descriptifs. On ne dispose donc pas d'études descriptives approfondies qui aient, par exemple, la couverture empirique de l'*Essai de grammaire de la langue française* de Damourette et Pichon (1911-1940). Ce travail d'élaboration de descriptions fines, approfondies et couvrantes est donc prioritaire (Abeillé, Godard & Miller, à paraître). Le problème d'INTÉGRATION, quant à lui, présente deux facettes. D'une part, les connaissances linguistiques sont partitionnées en *niveaux*, plus ou moins autonomes qu'il est essentiel de pouvoir interfacer les uns aux autres de façon efficace. D'autre part, même à l'intérieur d'un niveau, les *représentations* utilisées peuvent être très différentes d'une perspective théorique à l'autre et les niveaux eux-mêmes peuvent varier. Ainsi, pour caractériser l'*ordre des constituants* à l'intérieur d'un syntagme, une théorie comme la théorie principes-paramètres classique (Chomsky 1981, PP) utilisera essentiellement trois paramètres binaires (tête initiale/finale, assignation de Cas vers la gauche/droite et assignation de θ-rôle vers la gauche/droite)[3],

---

simplifiée du DiCo (Mel'cuk et al. 1995 : 224) : action sociale : personne X ~ Y à personne Z pour argent W.
Les termes *action sociale*, *personne* et *argent* sont des types contraignant respectivement le prédicat lui-même (*action sociale*), le premier et le troisième argument (*personne X* et *personne Z*) ainsi que le quatrième argument (*argent W*).

2 Les travaux des différentes équipes travaillant sur des *lexiques-grammaires* (Gross, 1975) constituent une exception d'autant plus notoire qu'elle semble isolée.

3 Cf. Travis (1989) et Fodor & Crain (1990). Les paramètres en question ont des effets ailleurs que dans le domaine des contraintes d'ordonnancement. Ceci constitue une qualité pour les tenants de la théorie PP, puisque cela ajoute à la «richesse de sa structure déductive». Du point de vue de la représentation des connaissances linguistiques, c'est plutôt un défaut, puisque cette propriété empêche de maintenir la transparence fonctionnelle des descriptions et donc la modularité des représentations. Ce(tte) défaut/qualité est maintenu(e) dans le programme minimaliste (Chomsky 1995; Epstein 1996 réd., PM).

tandis que des théories basées sur l'unification[4], comme la grammaire syntagmatique généralisée (Gazdar et al. 1985, GSG) ou la grammaire syntagmatique endocentrique (Pollard & Sag 1994, GSE) utilisent un type particulier de règles/schémas, les PL-règles/schémas.

En français et en japonais, les sujets précèdent en général le syntagme verbal central de la proposition, alors qu'en malgache les sujets suivent le syntagme (verbal ou non) central de la proposition. En français et en malgache, les compléments non pronominaux apparaissent après la tête lexicale (verbe, nom, adjectif ou préposition) dont ils dépendent, alors qu'en japonais, ils apparaissent avant[5].

(1) (a) Je suis allé à la montagne avec mon frère hier.

(b) [P [P [SN *je* SN] [SV [V *suis allé* V] [SP [Prép *à* Prép] [SN [Dét *la* Dét] [N *montagne* N] SN] SP] [SP [Prép *avec* Prép] [SN [Dét *mon* Dét] [N *frère* N] SN] SP] SV] P] [SP_{Adv} [Adv *hier* Adv] SP_{Adv}] P]

(2) (a) Nandeha tany an-tendrombohitra niaraka tamin'ny rahalahiko aho omaly.
$N_{PASSÉ}$-aller $t_{PASSÉ}$-LOC LOC-montagne $t_{PASSÉ}$-être ensemble $t_{PASSÉ}$-avec Dét frère-POSS1ps PRO1ps hier

(b) [P [P [SV [SV [V *nandeha* V] [SP [Prép *tany* Prép] [SN [N an-tendrombohitra N] SN] SP] SV] [SV [V *niaraka* V] [SP [Prép *amin* Prép] [SN [Dét *ny* Dét] [N *rahalahiko* N] SN] SP] SV] SV] [SN *aho* SN] P] [SP_{Adv} [Adv *omaly* SP_{Adv}] P]

(3) (a) Kinô boku wa nii-san to yama ni nobori-mashi-ta.
Hier moi TOP frère-HON COM montagne DAT monter-AUX-PASSÉ

(b) [P [SP_{Adv} [Adv Kinô Adv] SP_{Adv}] [P [SN [N *Boku* N] [Postp *wa* Postp] SN] [SV [SN [N *niisan* N] [Postp *to* Postp] SN] [SN [N *yama* N] [Postp *ni* Postp] SN] [V *noborimashita* V] SV] P] P]

---

4 Cf. Abeillé (1993).

5 Pour des raisons de simplicité, nous avons choisi une représentation par crochets étiquetés indentés de la structure syntagmatique. Cette représentation utilise des étiquettes qui sont des abréviations des catégories utilisées en GSG/GSE plutôt que de celles utilisées en PP. Une représentation de la structure syntagmatique dans le cadre PP serait nettement plus complexe que celle ci-dessus, mettant en jeu plusieurs niveaux (D-structure, S-structure, etc.), donc plusieurs représentations distinctes, chacun de ces niveaux contiendrait une multitude de projections fonctionnelles (SC, SI, SD) avec une grande quantité de catégories vides (INFL, COMP, DÉT, etc.), tout en n'étant pas nécessairement plus informative. Pour ce qui nous concerne ici (la distribution des catégories lexicales effectivement réalisées relativement aux catégories syntagmatiques), une représentation de S-structure dans le cadre PP, dépouillée de tous ces artifices (i.e., projections fonctionnelles, catégories vides et traces de mouvements) serait essentiellement congruente à la représentation donnée ici

Dans le cadre PP, on pourrait poser

(a) que le français et le japonais ont un sujet extrait en SS, alors que le malgache aurait un sujet extrait en FL et

(b) que le français et le malgache sont des langues à tête initiale, avec assignation de Cas et de θ-rôle vers la droite, alors que le japonais est à tête finale, avec assignation de Cas et de θ-rôle vers la gauche[6].

(4) Paramètres PP

(a) Français

SUJET	TÊTE	CAS	Θ-RÔLE
SS	**Initiale**	**Droite**	**Droite**

(b) Malgache

SUJET	TÊTE	CAS	Θ-RÔLE
FL	**Initiale**	**Droite**	**Droite**

(c) Japonais

SUJET	TÊTE	CAS	Θ-RÔLE
SS	**Finale**	**Gauche**	**Gauche**

Dans le cadre GSG/GSE, on pourrait poser que le SUJET est initial en français et en japonais et final en malgache et que ce sont les CATÉGORIES LEXICALES (têtes, mais également spécificateurs — déterminants, auxiliaires, modificateurs, etc.—) qui doivent apparaître avant les catégories syntagmatiques en français et en malgache, alors qu'elles apparaissent après ces dernières en japonais.

---

6 Il ne s'agit pas ici d'une analyse mais simplement d'une illustration. En fait, la description des contraintes d'ordre des mots en français, en malgache et en japonais dans le cadre PP serait beaucoup plus complexe que ne le laissent soupçonner ces tableaux. L'extraction devrait être contrainte de façon explicite. La valeur des quatre paramètres serait également paramétrisable. En français, par exemple, l'assignation du Cas accusatif par V ou oblique par Prép doit se faire vers la droite, mais l'assignation du Cas nominatif par INFL, vers la gauche. De plus, certaines têtes devraient être initiales, d'autres finales. On notera également que ce système rend nécessaire une multiplication de niveaux, puisqu'il exclut a priori des structures «plates» où la tête est enrobée de part et d'autres de constituants (spécificateurs d'un côté et compléments de l'autre, par exemple [$_{SN}$ Dét N SP] ou [$_{SV}$ Aux V SN] en français). Une telle structure minimale ne serait disponible que pour une langue où le paramètre de tête (initiale/finale) ne serait pas actif.

(5) (a) Français, japonais

PL : [SUJET(α)] ≪ α

(b) Malgache

PL : α ≪ [SUJET(α)]

(6) (a) Français, malgache

PL : [NIVEAU : **lexical**] ≪ ¬[NIVEAU : **lexical**]

(b) Japonais

PL : ¬[NIVEAU : **lexical**] ≪ [NIVEAU : **lexical**]

Si l'on tentait de calquer en partie la théorie PP en GSG/GSE, on pourrait postuler un trait [TÊTE : α] caractérisant les têtes de syntagmes[7] et avoir des PL-règles comme en (7).

(7) (a) Français

PL : [TÊTE : α] ≪ ¬[TÊTE : α]

(b) Japonais

PL : ¬[TÊTE : α] ≪ [TÊTE : α]

[7] En fait, Zwicky (1988) postule quelque chose de semblable. Dans le système de Gazdar et al. (1985), ce trait serait, pour les têtes lexicales, une abréviation de [SOUS-CAT : n], où n est l'index d'une règle. Quant aux têtes non lexicales, si elles peuvent porter le trait [SOUS-CAT : n], comme le proposent Pollard et Sag (1987), elles sont identifiables de façon analogue. Sinon, il faudrait construire un trait ad hoc. Cependant, on remarque que leur comportement est différent de celui des têtes lexicales. Par exemple, les têtes lexicales précèdent leurs compléments, mais SV suit son sujet en français. C'est d'ailleurs ce qui nous amène à exprimer les contraintes d'ordre en termes de NIVEAU (lexical ou non lexical) plutôt que de fonction (tête vs spécificateurs ou compléments).

Si, par exemple, on cherche à intégrer une description du SN (français ou japonais) suivant le cadre GSG/GSE et une description de la proposition conforme au cadre PP, il faudra d'abord opérer les coercitions nécessaires dans un sens (PARAMÈTRES$_{TÊTE/CAS/\Theta}$ > PL-RÈGLES) ou dans l'autre (PL-RÈGLES > PARAMÈTRES$_{TÊTE/CAS/\Theta}$). Une telle opération ne pose pas de problème de principe, mais exige un investissement considérable. On pourrait penser que le problème de l'intégration des REPRÉSENTATIONS peut être évité en choisissant une fois pour toutes une théorie et en s'y tenant[8]. L'expérience montre que tel n'est pas le cas[9]. La description de fragments importants de plusieurs langues naturelles présuppose la possibilité de réinterpréter les représentations de diverses théories dans la théorie choisie. De plus, même à l'intérieur d'une théorie donnée, la dynamique de la description entraîne des changements non négligeables de la théorie sous-jacente. Pourquoi dès lors ne pas faciliter directement cette réinterprétation? C'est précisément ce qu'une analyse en termes de DOMAINES, comme celle esquissée ci-haut pour le domaine des contraintes d'ordre permet de faire. Les institutions grammaticales sont les recueils de ces différents domaines, qui peuvent aussi bien être définis a priori qu'enrichis par l'ajout de nouvelles descriptions. Ainsi, pour la description syntaxique générale, on posera les domaines suivants :

8 Quant au problème de l'intégration des NIVEAUX, il demeure, de toute façon, incontournable.

9 Jusqu'à tout récemment, le groupe EUROTRA semblait favoriser un tel type d'approche. Cf. Arnold & des Tombe (1987).

Structure syntagmatique

CATÉGORIES
ORDONNANCEMENT
HIÉRARCHIE

Dépendances

GOUVERNEMENT	LIAGES
sélection	anaphore
rection	coréférence
quantification	
accord	
FILTRAGES	LIAGES-FILTRAGES
répulsion	contrôle
absorption	"mouvements"[10]
haplologie	
anaccord	
exclusions	

BORNAGES

Globaux/locaux Horizontaux/verticaux

Figure 1 : Domaines syntaxiques

Cette organisation est très souple. Si, par exemple, on voulait introduire la notion de *rayon*[11] de Damourette et Pichon dans une description formalisée, elle trouverait tout de suite sa place distribuée dans les domaines des CATÉGORIES et de la rection et l'on pourrait immédiatement voir qu'elle correspond aux valeurs du trait TFORME, telles qu'elles apparaissent comme valeurs de SOUS-CAT dans GSE. Si même l'on voulait introduire une notion ne correspondant à aucun des domaines prédéfinis, comme la

10 On regroupe sous ce terme inspiré de la métaphore transformationnelle, les liages-filtrages caractérisés en PP/PM comme A, $\bar{A}$, A ou parasites. En GSG/GSE, ou GLF il s'agit non pas de liages entre position pleine et vide, mais entre structures informationnelles, qui peuvent ou non correspondre à un noeud spécifique d'un arbre syntagmatique.

11 RAYON : classification des compléments selon leur introduction : «Je vois *mon père*» — direct; «Il a couru *vers la porte*» — *vers*; «Priez *pour moi*» — *pour*; «Arriver *à la ville*» — *à*; «Le curé *de Bazeille*» — *de*; «Je sors *le matin*» hors —.
Damourette, J. et E. Pichon (1991-1940) : *Des mots à la pensée. Essai de grammaire de la langue française. Compléments*, 13, D'Artrey, Paris, 1971.
(Le trait — doit être lu *rayon*. J'ai corrigé certaines marques typographiques incohérentes. JYM)

visée[12], toujours empruntée à Damourette et Pichon, on pourrait créer un nouveau type de liage correspondant et s'apercevoir que ce liage correspond à la notion de *première paire d'éléments dans la liste d'oblicité* ARGS en GSE.

On remarquera que ces domaines syntaxiques n'incluent pas de domaines correspondant directement aux notions configurationnelles, si importantes en PP/PM. On pourrait soit introduire ces notions configurationnelles comme portant sur de nouveaux sous-domaines des BORNAGES[13], soit les faire dériver entièrement d'autres notions non configurationnelles. Comme les notions configurationnelles sont omniprésentes dans toutes les versions de PP/PM[14], elles affectent à peu près tous les domaines (même l'information grammaticale morphologique y est codée configurationnellement). Il semble donc difficile de la factoriser. De plus, le rôle filtrant de ces notions est extrêmement faible : dans tout arbre non dégénéré (i.e., tout arbre branchant) *tous* les noeuds, sauf la racine, sont dans la relation de c-commande, soit comme c-commandeur, soit comme c-commandé. Aussi avons-nous opté plutôt pour la réduction systématique à des notions non configurationnelles (Morin 1996).

REPRÉSENTATION DES CONNAISSANCES LINGUISTIQUES : GRAMMAIRES D'OBJETS ET DE CONTRAINTES

Classiquement, il existe deux grands types de représentation des connaissances, qu'il s'agisse de connaissances linguistiques ou non : les REPRÉSENTATIONS PAR OBJETS et les REPRÉSENTATIONS PAR CONTRAINTES. Les REPRÉSENTATIONS PAR OBJETS définissent des univers d'objets, dotés de propriétés, entrant dans des relations et, éventuellement, capables d'actions (acteurs). Ces représentations mettent l'accent sur l'autonomie des objets, sur leur regroupement en classes organisées hiérarchiquement. Les objets d'une hiérarchie de classes partagent les uns avec les autres des propriétés héritées de leurs classes ancestrales communes. Ils peuvent prendre des valeurs par défaut ou déléguer à leurs ancêtres différentes fonctions. Cette conception semble intuitivement correspondre plus directement à notre ontologie naïve. Les REPRÉSENTATIONS PAR CONTRAINTES définissent des univers de relations abstraites ou contraintes. Les individus ou objets n'y existent et n'y sont identifiables qu'à travers les contraintes qu'ils satisfont. Ce type de représentation favorise l'abstraction et l'inférence. Cette conception semble correspondre à une ontologie plus formelle. Au niveau linguistique, on peut dire qu'un DICTIONNAIRE constitue une REPRÉSENTATION PAR OBJETS du langage, alors qu'une GRAMMAIRE en constitue plutôt une REPRÉSENTATION PAR CONTRAINTES.

---

12 VISÉE : rapport établi par un verbe entre son sujet et son about, c'est-à-dire celui de ses compléments auquel il tend par la force de sa signification (objet, attribut, etc.).
Damourette, J. et E. Pichon (1991-1940) : *Des mots à la pensée. Essai de grammaire de la langue française. Compléments*, 13, D'Artrey, Paris, 1971.

13 Les contraintes configurationnelles, comme *c-commande, gouvernement*, etc., ont un rôle essentiellement filtrant et non pas constructif, comme on l'a montré dans Morin (1996).

14 Sauf peut-être la réduction de c-commande à des notions transdérivationnelles d'Epstein (1996, réd.). Mais cette dernière introduit tellement de complexité cachée qu'il s'agit purement d'un exercice académique.

Objets

Représentation à base d'objets : grammaires à base lexicale

Plusieurs théories grammaticales récentes[15] partent de l'hypothèse que les connaissances grammaticales (du moins celles qui sont spécifiques à une langue) sont essentiellement rassemblées dans le lexique. Cependant, outre un certain nombre de remarques et de travaux intéressants, mais de nature plutôt programmatique, sur le rôle et le fonctionnement du lexique, sur sa position relative face aux autres composantes, sur sa structure générale ou sur la forme des entrées lexicales, très peu de chercheurs se sont penchés sur la construction effective de lexiques explicites[16]. Dans les théories linguistiques basées sur la notion d'*information grammaticale*, le rôle du lexique est particulièrement important, puisque celui-ci y constitue le répertoire principal (tant par l'envergure que par la complexité) de l'information grammaticale spécifique à une langue. De plus, dès que l'on se préoccupe de la *couverture* d'une description, tout autant que de sa rigueur, les problèmes associés à la construction de vastes lexiques deviennent incontournables[17].

En traitement des langues naturelles, d'autres facteurs encore contribuent à donner au lexique un rôle primordial. Le lexique contient l'ensemble des objets linguistiques directement accessibles et donc de coût constant, en termes de ressources. Les autres objets (structures, relations) ne sont accessibles que par construction[18] et sont donc de coût variable (éventuellement prohibitif) toujours en termes de ressources. En gros, il est plus facile de vérifier une information que de la (re)construire. Ce fait est reconnu, au moins implicitement, par la plupart des chercheurs du domaine.

À un niveau abstrait, un lexique constitue une *représentation par objets* des connaissances grammaticales. C'est un ensemble d'objets autonomes appartenant à des classes organisées en des réseaux plus ou moins serrés (e.g. catégories, catégories lexicales, catégories lexicales majeures ou verbes, verbes transitifs, verbes transitifs à double objet, etc.)[19]. Les objets des sous-classes héritent des propriétés de leurs superclasses (ou leur délèguent des propriétés). L'information sur le comportement syntaxique d'un objet (par exemple, la transitivité d'un verbe) est partiellement stockée dans la classe de l'objet et n'a pas à être répétée complètement dans l'entrée lexicale de celui-ci. Si un verbe appartient à la classe des verbes transitifs, il aura un comportement syntaxique de verbe transitif, dont les détails sont fixés une fois pour toutes et n'ont pas à être répétés. Seules les propriétés plus idiosyncratiques auront à être spécifiées dans

---

15 Par exemple PP, où lexique et paramètres sont fondamentaux, la grammaire lexicale-fonctionnelle (GLF) ou GSE.

16 Notons cependant les travaux de Gross (1975) sur les *lexiques-grammaires* et ceux de Mel'cuk (1984, 1988, 1992) sur les *dictionnaires explicatifs et combinatoires* (DEC) qui, malgré leur intérêt, s'intègrent difficilement aux théories syntaxiques issues de la grammaire générative, qui nous intéressent plus particulièrement ici.

17 Sur le problème de la couverture, cf. Bouchard, Emirkanian et Morin (1991).

18 Si l'on compare une analyse à une démonstration, les entrées lexicales constituent autant d'axiomes, directement démontrés, alors que les structures ne peuvent être démontrées que de façon constructive, à partir des règles du système, ce qui constitue une source de complexité calculatoire.

19 GSE étend cette structure de réseau en des hiérarchies de TYPES universelles.

l'entrée lexicale même (par exemple, le fait de régir un cas particulier pour l'objet direct). Ce type de représentation permet de construire des descriptions de façon très modulaire, incrémentale et dynamique et d'exprimer des valeurs par défaut. Ainsi, une propriété qui peut sembler idiosyncratique, comme le fait d'avoir une forme médio-passive en *se* sera d'abord notée à l'intérieur des entrées lexicales spécifiques (*vendre, laver*, etc.). Par la suite, elle pourra migrer hors de ces entrées et former une nouvelle classe où seront rassemblées les propriétés générales des médio-passifs, les entrées ne comportant plus qu'une référence à une fonction générique *médio-passif* indiquant que le verbe en question est dans le domaine de cette fonction.

En fait, toute la connaissance grammaticale peut être représentée en termes d'objets. Non seulement les objets lexicaux sont associés à des règles (ou des familles de règles[20]), mais les règles elles-mêmes peuvent être vues comme des objets abstraits. Les règles et les principes sont organisés en réseaux de classes et de sous-classes (correspondant en partie aux «modules» d'une théorie comme PP ou GSG/GSE). Elles ont des propriétés qui peuvent être héritées (ou déléguées). Ainsi, une règle ou un principe en PP a différents attributs : une strate ou un ensemble de strates d'application (DS, SS, FL, etc.), des paramètres, etc. Par exemple, l'assignation de CAS en PP peut être vue comme une classe d'objets ayant des paramètres de CAS, (eux-mêmes formés d'une VALEUR (par défaut *acc*) et d'un ASSIGNATEUR), d'une DIRECTION et d'une STRATE d'application (qui serait, par défaut la *S-tructure*).

(8) Assignation de CAS (PP)

$$\left[\begin{array}{ll} \text{CAS} : & \left[\begin{array}{ll} \text{VALEUR} : & \textbf{\textit{c}} \ (\textbf{\textit{acc}}) \\ \text{ASSIGNATEUR} : & \left\{\begin{array}{l} \textbf{\textit{V,}} \\ \textbf{\textit{Infl,}} \\ \textbf{\textit{Prép,}} \\ \ldots \end{array}\right\} \end{array}\right] \\ \text{DIRECTION} : & \left\{\begin{array}{l} \textbf{gauche} \\ \textbf{droite} \end{array}\right\} \\ \text{STRATE} : & \textbf{\textit{s}} \ (\textbf{\textit{S-structure}}) \end{array}\right]$$

De même, en GSG/GSE, une DI-règle/clause du DI-principe a une IDENTITÉ formée d'un NOM ET d'un INDEX ET d'une FAMILLE, une FONCTION —décomposition, adjonction, coordination—, et un CONTENU, CONSTITUÉ d'une MÈRE (ou PARTIE GAUCHE), d'un ensemble de FILLES (ou PARTIE DROITE), qui peuvent être des FILLES-TÊTES (dont le NIVEAU est, par défaut, lexical), des FILLES-COMPLÉMENTS, des FILLES-SPÉCIFICATEURS ou des FILLES-ADJOINTES. La présence de tous ces attributs est héritée, pour chaque DI-règle, du fait même qu'il s'agisse d'une DI-règle.

---

[20] Cf. Morin (1989), Blache (1990), Blache & Morin (1990) sur différentes notions de *famille de règles*.

$$\left[\begin{array}{ll} \text{IDENTITÉ :} & \left[\begin{array}{l}\text{NOM : } \quad n \\ \text{INDEX : } \quad i \\ \text{FAMILLE : } f\end{array}\right] \\ \text{FONCTION :} & \left\{\begin{array}{c}\textbf{décomposition} \\ \textbf{adjonction} \\ \textbf{coordination}\end{array}\right\} \\ \text{CONTENU :} & \left[\begin{array}{ll}\text{MÈRE} & p \\ \text{FILLES :} & \left[\begin{array}{l}\text{FILLES-TÊTES : } t \text{ ([NIVEAU : } \textbf{lexical]}) \\ \text{FILLES-COMPLÉMENTS : } \textbf{c} \\ \text{FILLES-SPÉCIFICATEURS: } \textbf{s} \\ \text{FILLES-ADJOINTES : } \quad \textbf{a}\end{array}\right]\end{array}\right]\end{array}\right]$$

Les GRAMMAIRES D'OBJETS constituent donc un cadre où différents types de connaissances grammaticales (correspondant, par exemple, à différentes théories linguistiques) peuvent être exprimés. L'élaboration de tels cadres de *représentation par objets* de la connaissance grammaticale (les GRAMMAIRES D'OBJETS) constituera un des objectifs fondamentaux de la recherche en génie linguistique dans les années à venir.

Contraintes

Représentations à base de contraintes : grammaires de règles et grammaires de principes

On oppose souvent les théories grammaticales *basées sur des règles* (comme GSG ou GLF) et les théories *basées sur des principes* (comme PP)[21]. En fait, les grammaires de principes, tout comme les grammaires de règles (ainsi que les grammaires traditionnelles d'ailleurs), constituent des *représentations par contraintes* des connaissances grammaticales. L'avantage de telles représentations est qu'elles permettent l'abstraction sur les objets et leur représentation. On peut exprimer des contraintes sans avoir à faire référence aux objets spécifiques qui leur seront soumis. Les contraintes s'appliquent à des variables qui n'ont pas à être instanciées, mais uniquement typées abstraitement[22].

Cependant, les représentations par contraintes posent de délicats problèmes de cohérence et de complexité. Comment comparer, par exemple, les hypothèses sur le

[21] En fait, cette opposition est plus une opposition méthodologique qu'une opposition conceptuelle. Méthodologie *ascendante* : du spécifique au général dans le cas des *grammaires de règles* vs méthodologie *descendante* : du général au spécifique, dans le cas des *grammaires de principes*. La notion de *principe* est malheureusement assez floue et sert parfois à masquer une absence de rigueur et une confusion entre spécifique (qui s'oppose à général) et explicite (qui s'oppose à implicite). La concision, la clarté ou la simplicité peuvent constituer des qualités d'une description, qui peut entraîner un peu d'implicite. Mais l'implicite n'est jamais une qualité en soi (du moins en science).

[22] D'ailleurs, toute contrainte constitue elle-même une fonction de typage abstrait.

liage[23] d'une théorie comme PP, et celles d'une théorie comme la GSG[24] ? Il faudrait pouvoir définir les théories en termes de contraintes sur certains *domaines* définissables indépendamment de la théorie. C'est ce qu'ont fait Miller (1990) pour les *systèmes de liage*, Gazdar et al. (1988) pour les *théories catégorielles*, Pullum et Barker (1990) pour les relations de *commande* et nous-même pour différents domaines syntaxiques (Morin 1988, 1989, 1996).

En traitement des langues naturelles, cette nécessité théorique s'associe à des exigences pratiques. Idéalement, on voudrait pouvoir pratiquer un certain éclectisme éclairé, qui permettrait d'importer dans un système **S** les contraintes exprimées par la théorie **A** sur un domaine **D**, de les combiner avec les contraintes exprimées par la théorie **B** sur un domaine **D'** (non nécessairement entièrement distinct de **D**) et avec les contraintes exprimées par la théorie **C** sur le domaine **D''** et ainsi de suite.

Au niveau syntaxique, on pourrait, par exemple, disposer

(a) d'une représentation relativement adéquate des phénomènes **DV** de DIATHÈSE VERBALE en français (les «voix» de la grammaire traditionnelle : ACTIVE, PASSIVE, PRONOMINALE, NEUTRE, MOYENNE, CAUSATIVE) exprimée dans le cadre de la *grammaire lexicale-fonctionnelle* **GLF**,

(b) d'une représentation des phénomènes de CLITICISATION PRONOMINALE **CP** dans le cadre de la théorie *gouvernement-liage* **PP**

et

(c) d'une représentation des phénomènes de COOCCURRENCE et d'ORDONNANCEMENT DES COMPLÉMENTS **COC** dans le cadre de la théorie de la *grammaire syntagmatique généralisée* **GSG**.

---

23 Ou plus précisément, sur le *liage-filtrage*. Sur cette notion de *liage-filtrage* (ensemble des liens globaux entre objets présents et objets «absents» —filtrés— dans la structure), cf. Morin (1989).

24 En PP, les phénomènes sont décrits au moyen de l'interaction d'un réseau de notions extrêmement complexe et mouvant d'un auteur à l'autre : *chaînes*, *gouvernement*, *K-marquage*, *θ-marquage*, *filtres*, *principe de projection*, *θ-critère*, A et ¬A-*positions*, etc. En GSG, on fait appel à des principes universels de propagation de valeurs (essentiellement le *principe de traits de pied* (Gazdar et al. 1985) ou le *principe de traits de liage* (Sag & Pollard 1987)) associés à un système de traits typés et à des restrictions intracatégorielles (restrictions de cooccurrence de traits, spécifications de traits par défaut) et extracatégorielles (DI-règles) sur les valeurs de certains traits. La comparaison de la théorie du liage PP et de celle de GSE est beaucoup plus facile, puisque la théorie du liage de GSE (Pollard & Sag 1994, ch. 6) est explicitement construite à partir de celle de PP, mais en utilisant la structure informationnelle et la hiérarchie d'oblicité (o-commande) plutôt que des notions configurationnelles (c-commande).

Dès lors, on pourrait vouloir intégrer ces trois fragments dans un système S :

(10)

S = s(**GLF(DV)**) & s(**PP(CP)**) & s(**GSG(COC)**)[25]

CONCLUSION

Nous avons examiné ici quelques-uns des problèmes liés à l'intégration des théories et des descriptions linguistiques en traitement des langues naturelles. Nous avons montré les avantages d'une approche par objets et contraintes et jeté les bases de ce que pourraient être des *institutions grammaticales*. La définition plus précise de ces domaines que constituent les institutions grammaticales, l'interprétation de fragments de descriptions dans ces termes et l'élaboration de grammaires d'objets et de contraintes constitueront les objectifs fondamentaux de notre recherche en ce qui concerne l'exploitation des théories et descriptions linguistiques en TAO.

RÉFÉRENCES

ABEILLÉ, Anne (1993) : *Les nouvelles syntaxes*, Paris, Armand Colin.

ABEILLÉ, Anne, GODARD, Danièle et Philip MILLER (à paraître) : *The Major Syntactic Structures of French*, Notes de cours d'ESSLLI-97, Aix-en-Provence.

ARNOLD, D. et L. des TOMBE (1987) : «Basic theory and methodology in EUROTRA», *in* Nirenburg, S. (1987 réd.) *Machine Translation, Theoretical and Methodological Issues*, Cambridge, Cambridge University Press, pp. 114-135.

BLACHE, Ph. (1990) : *L'analyse syntaxique dans le cadre des grammaires syntagmatiques généralisées, interprétations et stratégies*, Thèse de doctorat inédite, Faculté des Sciences de Luminy, Université d'Aix-Marseille II.

BLACHE, Ph. et J.-Y. MORIN (1990) : «Bottom-Up Filtering : A Parsing Strategy for ID/LP Grammars», *in* Karlsson, F. (1990 réd.) *COLING-90, Proceedings*, Helsinki, Université d'Helsinki et New-York, Association for Computational Linguistics.

BOUCHARD, L., EMIRKANIAN, L. et J.-Y. MORIN (1991) : «Détermination de la couverture dans les interfaces en langue naturelle», *in* Gauthier, G. (1991 réd.) *ICO-91*.

CHOMSKY, N. (1981) : *Lectures on Governement and Binding*, Dordrecht, Foris.

CHOMSKY, N. (1995) : *The Minimalist Program*, Cambridge, Mass., MIT Press.

DA SYLVA, L. (1990) : *Un parseur inspiré de la théorie gouvernement-liage*, Mémoire de maîtrise inédit, Département de linguistique et de traduction, Université de Montréal.

[25] s est ici une fonction générique d'interprétation de l'ensemble **t(d)** des énoncés de la théorie **t** sur le domaine **d**.

DAMOURETTE, Jacques et Édouard PICHON (1911-1940) : *Des mots à la pensée. Essai de grammaire de la langue française*, 7 volumes et un volume d'annexes, Paris, d'Artrey.

EPSTEIN, S. et al. (1996 réd.) : *Minimal Ideas*, Amsterdam, John Benjamins.

FODOR, J. & S. CRAIN (1990) : «Phrase Structure Parameters», *Linguistics and Philosophy*, 13, 6, pp. 619-659.

GAZDAR, G., KLEIN, E. PULLUM, G. et I. SAG (1985) : *Generalized Phrase Structure Grammar*, Oxford, Blackwell et Cambridge, MA, Harvard Univ. Press.

GAZDAR, G., PULLUM, G. CARPENTER, R. KLEIN, E., HUKARI, T. & R. LEVINE (1988) : «Category structures», *Computational Linguistics*, 14, 1, 1-19. (Traduction française dans Torris, T. et Ph. Miller (1990 réd.) *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Paris, Hermès).

GOGUEN, J.A. et R.M. BURSTALL (1983) : «Introducing Institutions», *in* Clarke, E. et D. Kozen (1983 réd.) *Logics of Programs*, Lecture Notes in Computer Science, 164, Berlin, Springer-Verlag, pp. 221-256.

GOGUEN, J.A. et R.M. BURSTALL (1985) : *Institutions : Abstract Model Theory for Computer Science*, Rapport CSLI-86-54, Stanford, Center for the Study of Language and Information, Stanford University.

GOGUEN, J.A. et R.M. BURSTALL (1986) : *A Study in the Foundations of Programming Methodology : Specifications, Institutions, Charters and Parchments*, Rapport CSLI-85-30, Stanford, Center for the Study of Language and Information, Stanford University.

GROSS, Maurice (1975) : *Méthodes en syntaxe*, Paris, Hermann.

JACKENDOFF, R. S. (1987) : *Consciousness and the Computational Mind*, Cambridge, MA, MIT Press.

MEL'CUK, I. et al. (1984, 1988, 1992) : *Dictionnaire explicatif et combinatoire du français, contemporain Recherches lexico-sémantiques* I-III, Montréal, Presses de l'Université de Montréal.

MILIÇEVIÇ, J. (1997) : *Étiquettes sémantiques dans un dictionnaire formalisé du type Dictionnaire Explicatif et Combinatoire*, Mémoire de maîtrise, Département de linguistique et de traduction, Université de Montréal.

MILLER, Ph. (1990) : «Systèmes de liage», *in* Torris, T. et Ph. Miller (1990 réd.) *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Paris, Hermès.

MILNER, J.-C. (1989) : *Introduction à une science du langage*, Paris, Seuil.

MORIN, J.-Y. (1988) : «Prédicats théoriques et données externes. Syntaxe diachronique», *Revue canadienne de linguistique*, numéro spécial, *Linguistic Theory and External Evidence*, 33, 4, pp. 443-475.

MORIN, J.-Y. (1989) : *Syntaxe*, Département de linguistique et philologie, Université de Montréal.

MORIN, J.-Y. (1996) : «Configuration vs. information. An informational explanation of command relations», *in* Park, B.-S. & J.-B. Kim (1996 réd.) *PACLIC-11, Language, Information and Computation*, Séoul, LERI, Kyung Hee University, pp. 11-20.

POLLARD, C.J. (1984) : *Generalized Phrase Structure Grammars, Head Grammars and Natural Language*, Thèse de Ph. D. inédite, Stanford University.

POLLARD, C.J. & I. SAG (1994) : *Head-Driven Phrase Structure Grammar*, Chicago, Univ. of Chicago Press.

PULLUM, G. et C. BARKER (1990) : «A Theory of Command Relations», *Natural Language and Linguistic Theory.*

SAG, I. et C.J. POLLARD (1987) : *Information-Based Syntax and Semantics, Vol. I : Fundamentals*, Stanford, CSLI et Chicago University of Chicago Press.

SAINT-DIZIER, P. (1989) : «Programming in logic with constraints for natural language processing», *ACL Europe 4*, pp. 87-94.

SAINT-DIZIER, P. & P. SZPAKOWICZ (1989 réd.) : *Logic Programming and Logic Grammars*, Londres, Ellis Horwood.

TRAVIS, L. (1989) : «Parameters of Phrase Structure», *in* Baltin, M. & A. Kroch (1989 réd.) *Alternative Conceptions of Phrase Structure*, Chicago, University of Chicago Press.

ZWICKY, A. (1986) : *Interfaces*, Colombus, Ohio State University Working Papers in Linguistics, # 32.

ZWICKY, A. (1988) : «Direct reference to Heads», *Folia Linguistica*, 22.3-4, pp. 397-404.

LA MÉMOIRE DES PARTICIPES PRÉSENT ET PASSÉ

Poul Søren KJÆRSGAARD

Universite d'Odense, Odense, Danemark

1. INTRODUCTION

La présente contribution a pour but de faire le point sur un projet de recherche[1] qui examine les participes présent et passé en danois.

D'un point de vue immanent, il s'agit de décrire leur structure (simple ou composée), leur(s) fonction(s) ainsi que des critères permettant de les distinguer des adjectifs. Il s'agira aussi d'expliquer les contraintes régissant leur usage.

D'un point de vue contrastif, le projet s'est assigné le but de décrire la distribution des participes dans des langues voisines (les langues germaniques), mais aussi de décrire leurs équivalents dans d'autres familles de langues, notamment les langues romanes. Cette description déboucherait, à terme, sur l'établissement de procédés et de règles de traduction entre ces langues.

Dans cet article, nous allons prendre, dans un premier temps, comme point de départ les participes présent et passé en français pour mettre en lumière des comportements qui s'expliquent, nous semble-t-il, en faisant référence à l'idée d'une mémoire des mots. Cette expression est bien entendu à comprendre au sens figuré. Par mémoire, nous entendons le fait que des comportements de certains mots, par exemple leur combinaison avec d'autres mots, trouvent une explication si l'on admet l'existence de traces ou de mémoires du comportement dans un état antérieur. Concrètement, certains comportements des participes présent et passé nous paraissent explicables si l'on suppose l'existence d'un lien entre ces formes et des formes verbales finies ou infinitives. Nous présupposons donc que les participes sont dérivés de ces autres mots.

Le second volet de l'article examinera les participes présent et passé en danois, cette langue choisie comme représentante des langues germaniques. Là, nous constatons également qu'une description de leur structure et de leur fonction conduit à l'idée d'une

[1] Le projet décrit ici constitue avec cinq autres projets un projet d'ensemble, *Udforskning af dansk ordforråd og grammatik* , UDOG [Recherches sur le vocabulaire et la grammaire danois] visant à décrire le vocabulaire et la grammaire (syntaxe) de la langue danoise contemporaine. Il est coordonné par *Center for sprogteknologi* (Centre for Language Technology) de Copenhague et a été subventionné par le CNRS danois sous contrat n° 15-9018 ainsi que par les universités des six équipes.

mémoire de ces mots. Puisqu'il s'agit d'une autre langue, ce sont évidemment d'autres mécanismes qui entrent en jeu. Nous allons voir que la structure des participes en danois et en français conduit à des problèmes de traduction. Ces problèmes, qu'il s'agisse d'un traducteur humain ou non, deviennent explicites, si l'on s'appuie sur l'idée de mémoire, c.-à-d. de relations et de traits morphosyntaxiques, qui, certes, sont neutralisés en surface, mais qui déterminent, partiellement au moins, la combinatoire des participes.

Enfin dans la troisième partie de l'article, nous allons passer en revue l'éventail des structures utilisées en français pour rendre les constructions danoises.

2. LES PARTICIPES FRANÇAIS

2.1 Le participe présent

Le participe présent s'emploie sous deux formes différentes, l'une appelée participe présent, l'autre adjectif (verbal)[2]. Sans entrer dans le détail — nous renvoyons aux descriptions des manuels de grammaire, p.ex. Arrivé et al. (1986 : 472-473), Riegel et al. (1997 : 340-341) et de l'article de Halmøy (1984, 52-57ss.), il existe des critères morphologiques, syntaxiques, et lexico-sémantiques, qui permettent de distinguer l'un et l'autre. Résumons-les brièvement :

a) Le participe maintient l'orthographe [*qu/gu*]-ant, alors que certains adjectifs (verbaux) choisissent l'orthographe [*c/g*]-ant. Le participe conserve aussi *-ant*, alors que certains adjectifs (verbaux) prennent *-ent*. En plus, le participe reste inchangé, alors que l'adjectif (verbal) s'accorde en genre et en nombre avec le nom auquel il se rapporte.

b) Le participe se combine avec les compléments en fonction de la valence et de la construction du verbe correspondant. Cette option n'existe pas pour l'adjectif (verbal). Le participe ne s'emploie pas comme attribut (du sujet), ce qui est le cas de l'adjectif (verbal)[3]. Les deux s'emploient comme épithète. Le participe ne se renforce pas par *très*, ce qui est le cas de l'adjectif (verbal) : *Un projet de loi (*très) intéressant les syndicats* vs *Un projet de loi (très) intéressant* (cité d'après Kjærsgaard, 1995 : 118). Épithète, le participe suivi de ses compléments est placé derrière le nom. Dans cette fonction et dans un style plutôt soutenu, l'adjectif (verbal) peut être antéposé au nom. Cette différence s'accompagne souvent d'une différence de sens, cf. ci-dessous.

---

[2] Dans un article intitulé *À propos de l'adjectif en -ant, dit "verbal"*, Odile Halmøy s'insurge contre cette étiquette. Elle propose, avec raison, de considérer la différence des deux formes comme une instance de la différence entre flexion et dérivation. Afin de marquer ce parti pris, nous avons adopté l'étiquette *adjectif (verbal)*.

[3] Halmøy (1994 : 53) observe que certains adjectifs (verbaux) entrent en collocation avec certains noms, de façon à ce que la fonction attributive entre eux cesse d'exister : *eau courante *→ cette est courante.*

c) Le participe des verbes perfectifs[4] dénote l'aspect inaccompli de l'action du verbe et la simultanéité avec le verbe fini de la phrase, alors que l'adjectif (verbal) décrit plutôt un état ou une propriété du nom auquel il se rapporte.
Le participe est formé du suffixe *-ant* accolé à la racine de l'infinitif, ce qui n'est pas toujours le cas de l'adjectif (verbal). C'est ainsi qu'on a en français le participe *connaissant*, p.ex. dans *des étudiants connaissant bien la langue française*. L'adjectif (verbal) correspondant, par contre, n'est pas ce lexème, mais *connaisseur*. On n'a donc pas **des étudiants connaissants de la langue française*, mais bien *des étudiants connaisseurs de la langue française*[5]. Par contre, on a en français l'adjectif (verbal) *reconnaissant*, p.ex. dans *la patrie reconnaissante*. Ce lexème s'emploie aussi comme participe, p.ex. *les Français reconnaissant le besoin de soutenir leur langue...*
À l'inverse, et pour des raisons inexpliquées, *reconnaissseur* est inexistant.
Notons enfin des différences de sens entre le participe et l'adjectif (verbal). Dans ces cas, c'est le participe qui conserve le sens de l'infinitif, alors que l'adjectif (verbal) change de sens. Halmøy (1984 : 59) donne l'exemple de *regardant*, avec le sens de *regarder* dans *un homme regardant la télé* vs *un homme regardant* où l'adjectif (verbal) prend la valeur de *avare*. D'autres exemples cités dans les manuels de grammaire sont *violant/violent, comptant/comptant* [*argent comptant*].

La relation syntaxique entre un nom et un participe/adjectif (verbal) est normalement celle existant entre le sujet et le verbe (fini) : *des projets intéressant l'ensemble des professeurs* vs *des projets intéressants*. Cette relation apparaît en paraphrasant l'expression à l'aide d'une proposition relative : *des projets qui intéressent l'ensemble des professeurs* vs *des projets qui intéressent/sont intéressants*[6]. La relation sémantique entre un nom et un participe/adjectif (verbal) est active ou passive selon le sens du verbe : *les chiens aboyants* vs *les nations naissantes*.

Épithète, l'adjectif (verbal) autorise parfois un relâchement des contraintes valencielles : *les villes environnantes*, alors qu'on ne trouverait guère le verbe *environner* sans complément d'objet direct.

C'est dans ce contexte qu'on trouve une première instance de mémoire, car l'adjectif (verbal) entretient parfois des relations diverses avec la tête du groupe nominal dans lequel il se trouve.

Exemples cités d'après Arrivé et al. (1984 : 473) et Riegel et al. (1997 : 341) :

une couleur voyante	*→ une couleur qu'on voit (très bien)*	sens passif
une personne méfiante	*→ une personne qui se méfie*	sens réfléchi

[4] Cette distinction ne s'applique pas aux verbes imperfectifs (cf. Halmøy, 1984 : 55). Un exemple en est le verbe *surveiller* : *des femmes surveillant le square* vs *des femmes surveillantes*.

[5] D'autres exemples sont présentés par Halmøy (1984 : 49) : *un site *enchantant/enchanteur, un calme *trompant/trompeur, un siège *inclinant/inclinable*. Dans ces cas, le participe serait p.ex. *un siège s'inclinant en arrière, un calme trompant l'entourage.*

[6] Cette paraphrase ne constitue pas une relation bijective, c.-à-d. elle ne s'applique pas toujours dans les deux sens comme nous l'avons vu ci-dessus : *un site qui enchante → un site *enchantant.*

un revêtement glissant	→ *un revêtement qui fait glisser*	sens factitif
une soirée dansante	→ *une soirée durant laquelle on danse*	sens temporel
une rue passante	→ *une rue où beaucoup de gens passent*	sens locatif.

Ces exemples nous font dire que les adjectifs (verbaux) sont susceptibles de recouvrir des valeurs neutralisées, mais non pas disparues, car elles demeurent à l'état virtuel et réapparaissent au paraphrasage.

2.2 Le participe passé

Le participe passé s'emploie, comme le participe présent, de deux façons différentes, soit comme tête d'un groupe adjectival sans compléments (équivalent de l'adjectif (verbal)), soit comme tête d'un groupe adjectival avec compléments (équivalent du participe). La principale différence entre elles est, comme pour le participe présent, la gamme réduite de fonctions du dernier qui ne s'emploie guère comme attribut (du sujet) : *des chambres réservées* vs *des chambres sont réservées*, mais *des chambres à eux*[7] *réservées* vs *des chambres sont *[à eux] réservées.*

Contrairement à son homologue, le participe passé connaît cependant des contraintes de combinatoire et de formation.

Épithète d'un groupe nominal, le participe n'est en principe compatible qu'avec une tête qui fonctionne comme le complément d'objet direct du verbe fini. Cette règle constitue la contrepartie logique de la règle employée pour le participe présent, cf. ci-dessus.

Cela explique qu'on a *les paquets expédiés* (quelqu'un a expédié les paquets), mais non pas **les étudiants expédiés* (quelqu'un a expédié les paquets aux étudiants).

La règle connaît au moins deux entorses : quelle que soit la fonction syntaxique du nom, le rôle sémantique de celui-ci doit être celui d'un patient ou d'un thème (selon la typologie et la taxinomie employées). Cela explique l'existence de *les enfants morts pour la patrie*. Dans cet exemple, la relation syntaxique entre le nom et le participe épithète est celle d'un sujet et d'un verbe, mais leur relation sémantique est celle qui existe entre un verbe et un patient. La relation sémantique prime donc la relation syntaxique.

Comme pour le participe présent, il existe aussi des collocations dans lesquelles les rapports normaux sont suspendus ou dans lesquels la thèse d'une relation fixe (verbe-patient) n'a pas de sens. Mentionnons (d'après Riegel, 1997 : 344) *le journal parlé, la presse écrite, l'eau salée.*

Pour ce dernier exemple, on remarque qu'il est passé complètement dans la catégorie des adjectifs, un critère valable étant sa compatibilité avec *très*[8].

---

[7] *à eux* s'analyse comme un complément d'attribution (objet datif).

[8] L'auteur de cet article a esquissé une série de tests permettant de distinguer adjectifs et participes déverbaux en danois : Kjærsgaard et le Fevre Jakobsen (1997, à paraître).

Faisant abstraction des contre-exemples, il semble légitime d'affirmer que la combinaison d'un nom et d'un participe passé épithète dans la grande majorité des cas est régie (et contrainte) par des relations qui se trouvent sur un niveau différent de l'analyse, celle de la phrase ou celle des relations verbe-compléments. De là à affirmer que le participe passé épithète se souvient de (ou a mémorisé des) comportements inhérents au verbe, il n'y a qu'un pas.

Le participe passé connaît aussi une autre contrainte importante à son emploi comme épithète. Alors que l'adjectif (verbal) s'emploie comme épithète des verbes monovalents (intransitifs) inergatifs[9], il n'en est rien du participe passé :

eau courante	**eau courue*
eau dormante	**eau dormie*
chien aboyant	**chien aboyé*

L'emploi épithétique existe pour les deux participes des verbes monovalents ergatifs :

les enfants mourants	*les enfants morts*
le caramel fondant	*le caramel fondu*

La conclusion qui s'impose est que le comportement syntaxique des participes épithètes, y compris leurs /in-/compatibilités avec la tête du groupe nominal, reflète un embryon de mémoire des propriétés valables pour les verbes finis. On peut exprimer la même conclusion en disant que les contraintes valencielles (nombre des actants, nature de leurs constituants, leurs fonctions syntaxiques, leurs rôles sémantiques et les contraintes de sélections qui y sont associées) qui régissent les rapports entre le verbe fini et ses actants (sujet + compléments) sont mémorisés par les participes. Enfin, c'est un savoir intrinsèque qui demeure à l'état virtuel, récupérable au besoin.

3. LES PARTICIPES DANOIS

3.1 Les participes simples

Les participes danois sont employés moins fréquemment que leurs homologues français. Cela tient à différentes raisons que la présente contribution n'examinera pas.

Un participe présent danois se traduit parfois par un participe passé français ou vice versa :

et foretagende	*une entreprise*
(art. indéf.neut.sg.-part.pr. substantivé)	
en tilbundsgående undersøgelse	*une enquête approfondie*

[9] Les expressions *inergatif* et *ergatif* s'emploient selon la terminologie de la grammaire relationnelle, développée par David Perlmutter et autres. Un verbe inergatif est un verbe (intransitif) à la voix active dont l'action part du sujet : *courir*. Un verbe ergatif (ou inaccusatif) est un verbe (intransitif) à la voix active dont l'action atteint ou vise le sujet : *périr*.

(art. indéf.com.sg.[10]-part.pr.-nom sg.)

en befærdet gade *une rue passante*
(art. indéf.com.sg.-part.pa.-nom sg.)

Ces différences lexicales sont considérées comme des cas de l'arbitraire du signifiant et ne seront pas non plus traitées dans le cadre présent.

Les relations sémantico-syntaxiques qui valent pour les participes français, sont valables grosso modo pour les participes danois.

Le rapport normal entre un nom et un participe présent/adjectif (verbal) épithète est celui du verbe et de son sujet. Parallèlement, la relation qui existe entre un nom et un participe passé épithète correspond à celle qu'on trouve entre le verbe et son objet direct. On voit donc en danois, grosso modo, les mêmes contraintes qu'en français.

Pour les monovalents inergatifs, le danois se comporte comme le français[11] : le participe présent existe, alors que ce n'est pas le cas du participe passé :

de løbende atleter *les athlètes qui courent*[12]
(art. déf.pl.-part.pr.:*courir*-nom pl.:*athlète*)

*de *løbne atleter* **les athlètes courus*
(art. déf.pl.-part.pa.:*courir*-nom pl.:*athlète*)

de knoklende landmænd *les paysans qui bossent*
(art. déf.-part.pr.:*bosser*-nom pl.:*paysan*)

*de *knoklede landmænd* **les paysans bossés*

Les monovalents ergatifs (inaccusatifs) connaissent les deux participes :

det synkende skib *le bateau qui fait naufrage*
(art. def.sg.neut.-part.pr.:*faire naufrage*-nom sg.:*bateau*)

det sunkne skib *le bateau qui a fait naufrage*
(art. def.sg.-part.pa.:*faire naufrage*-nom sg.:*bateau*)

de sammenstyrtende huse *les maisons qui s'écroulent*
(art. def.pl.-part.pr.:*s'écrouler*-nom pl.:*maison*)

---

10 Le danois possède deux genres : le neutre et le commun (amalgame du masculin et du féminin).

11 Les langues française et danoise ne sont pas les seules à se comporter ainsi. À ma connaissance, la grammaire relationnelle pose cette relation comme universelle.

12 Les traductions fournies sont celles qui paraissent le plus appropriées pour une traduction hors contexte. C'est ainsi qu'un participe présent danois équivaut souvent à une relative française. Le contenu des parenthèses indiquent la partie du discours des mots danois.

de sammenstyrtede huse *le maisons écroulées/qui se sont écroulées*
(art. def.pl.-part.pa.:*s'écrouler*-nom pl.:*maison*)

Dans ces cas, la différence demeure partout celle qui existe entre une acception aspectuelle inaccomplie (simultanée) et une autre accomplie.

Pour les verbes bivalents inergatifs, on observe les patterns suivants :

de spisende børn[13] *les enfants qui mangent*
(art. déf.pl.-part.pr.:*manger*-nom pl.:*enfant*)

de spiste grønsager *les haricots consommés*
(art. déf.pl.-part.pa.:*consommer*-nom pl.:*haricot*)

de skabende erhverv *les métiers créateurs*[14]
(art. déf.pl.-part.pr.:*créer*-nom pl.:*métier*)

de skabte værdier *les plus-values créées*
(art. déf.pl.-part.pa.:*créer*-nom pl:*valeur*)

Dans aucun exemple cité, il ne serait possible de substituer un participe présent à un participe passé ou vice versa. Cela s'explique par des contraintes valencielles au niveau sémantique des verbes en question. Les verbes *spise* (manger, consommer) et *skabe* (créer) supposent normalement deux actants, dont le premier est sujet et animé et le second est objet direct et inanimé.

Le même pattern s'applique aux verbes ergatifs :

de lidende børn *les enfants qui souffrent*
(art. déf.pl.-part.pr.:*souffrir*-nom pl.:*enfant*)

de lidte tab *les pertes subies*
(art. déf.pl.-part.pa.:*souffrir*-nom pl.:*perte*)

---

13 Dans les exemples suivants, un actant, facultatif, est omis. Pour exprimer un nom d'agent, le danois peut même omettre tous les actants. C'est ainsi qu'on a *de spisende* (participe présent substantivé au plur.déf.) qui correspond à *les mangeurs*. Il existe une concurrence entre le nom d'agent et le participe présent substantivé, apparemment en danois comme en français : *de levende* vs *les vivants* (tous deux des participes); *vinderne* (plur.déf.) vs *les gagnants* (nom d'agent vs participe présent). Aucune règle ne permet de prédire si c'est l'une ou l'autre construction qui l'emporte (l'arbitraire du signifiant).
Pour les verbes trivalents, assez rares, le nom d'agent l'emporte sur le participe présent substantivé : *giverne* (nom-art.déf.pl. accolé au nom) - *les donneurs; lånerne - les emprunteurs/les prêteurs.*

14 La différence entre un participe présent tête d'un groupe à compléments et un adjectif (verbal) ne se manifeste pas au niveau lexical comme c'est parfois le cas en français, cf. par. 2.1.

3.2 Les participes composés

Les langues germaniques, dont le danois, se caractérisent encore par un phénomène quasiment inconnu des langues romanes : la faculté de préfixation d'un lexème (nom, adjectif, adverbe ou particule), le nouvel ensemble constituant un composé synthétique et assumant les mêmes fonctions syntaxiques que le participe simple, c.-à-d. principalement comme épithète d'un nom tête d'un groupe nominal ou comme attribut[15]. Il s'agit d'un processus de formation lexicale très productif en danois contemporain.

Cette construction synthétique existe aussi pour les noms et les autres parties du discours. Pour des raisons qui ne seront expliquées ici, il ne peut être préfixé qu'un seul lexème au participe (ou, comme nous l'appellerons dorénavant : participe déverbal (pour les distinguer des adjectifs)). Cela signifie que, pour les verbes trivalents, au moins un actant ne peut être exprimé.

Pour les participes présent et passé, le lexème préfixé et le nom tête du groupe nominal auquel se rapporte le participe déverbal, assument des rapports qui se résument par deux patterns archétypiques :

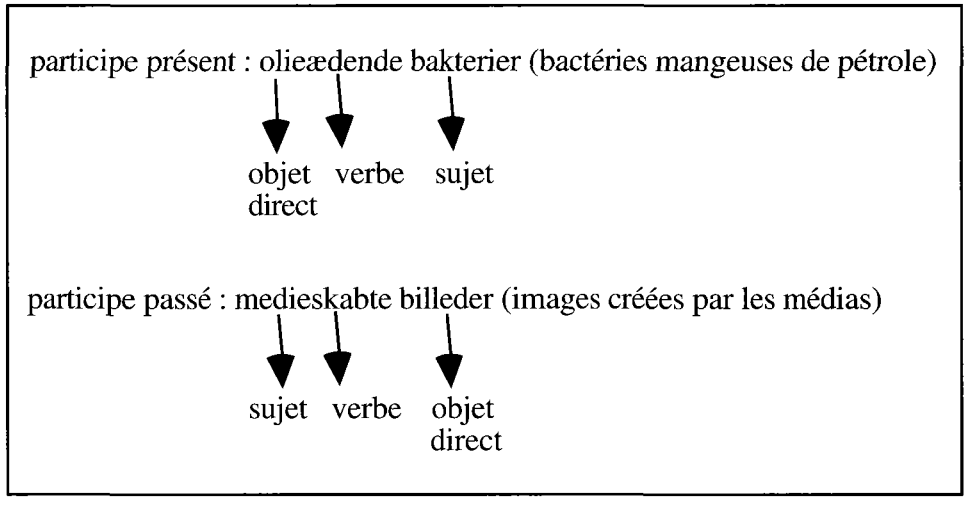


Figure 1

15 Les exceptions des langues romanes sont peu nombreuses et se répartissent en deux groupes :
a) Mots d'emprunt d'origine grecque ou latine qui correspondent aux participes danois (et a fortiori germaniques) :
cancérogène vs *kræftfremkaldende* [nom:*cancer*-part.pr.:*provoquer*];
bactéricide vs *bakterieædende* [nom:*bactérie*-part.pr.:*manger*].
b) Hapax, que je sache, en espagnol : *hispanohablante* vs *spansktalende*.
Ajoutons enfin que tel participe composé danois (ou germanique) correspond parfois à un lexème non composé en français : *computerstyret* [nom:*ordinateur*-part.pa.:*diriger/exécuter*] vs *informatisé*.

Ce sont ces patterns qui permettent de prédire la possibilité des constructions suivantes :

laboratoriekontrollerede fødevarer *des aliments contrôlés par des laboratoires*
([nom:*laboratoire*-part. pa.:*contrôler*]-nom pl.:*aliment*)

fødevarekontrollerende laboratorier *des laboratoires contrôlant les aliments*
([nom:*aliment*-part.pr.:*contrôler*]-nom pl.:*laboratoire*)

Ce sont aussi ces patterns qui permettent de prédire l'impossibilité des constructions inverses :

**fødevarekontrollerede laboratorier* **des laboratoires contrôlés par des aliments*
(nom:*aliment*-part.pa.:*contrôler*]-nom pl.:*laboratoire*)

**laboratoriekontrollerende fødevarer* **des aliments contrôlant les laboratoires*
([nom:*laboratoire*-part.pr.:*contrôler*]-nom pl.:*aliment*)

Les contraintes de sélection sémantiques du verbe *kontrollere* [contrôler] interdisent tout simplement ces constructions. Voici un nouvel élément qui, nous semble-t-il, justifie l'affirmation que le participe déverbal, dérivé du verbe, conserve un embryon de mémoire.

Aux fonctions syntaxiques indiquées s'ajoutent évidemment des rôles sémantiques : le sujet correspondant le plus souvent à l'agent et l'objet direct au patient/thème.

Le schéma reproduit ci-dessus est pourtant loin de représenter l'intégralité de la réalité linguistique. Si le participe déverbal dérive d'un verbe régissant un complément indirect (objet prépositionnel), la préposition est supprimée pendant le processus de préfixation, et structurellement, il n'y a donc pas moyen de voir de différence entre des objets direct et indirect préfixés. Les compléments circonstanciels, souvent exprimés à l'aide d'un groupe prépositionnel, se voient aussi privés de préposition : Il s'ensuit qu'on ignore la fonction et le rôle du préfixe (cas à distinguer : sujet ou complément circonstanciel accolé au participe passé; objet direct ou complément circonstanciel accolé au participe présent). Par conséquent, les constructions préfixées sont souvent structurellement ambiguës (voir exemple ci-dessous).

Puisque la plupart des participes déverbaux ne sont pas (et ne devraient pas être) lexicalisés dans les dictionnaires (ni pour humains, ni pour ordinateurs), il s'ensuit que ces ambiguïtés ne sont pas directement décelables, et que la détermination des relations entre le participe déverbal d'un côté, et l'élément préfixé et le nom tête du groupe nominal de l'autre, requièrent une analyse préalable. L'exemple suivant illustre cette ambiguïté :

vandforurenende virksomheder *des entreprises polluant l'eau*
([nom: *eau*-part.pr.:*polluer*]-nom pl.:*entreprise*)

olieforurenende virksomheder *des entreprises polluant <u>par</u> le pétrole*
([nom: *pétrole*-part.pr.:*polluer*]-nom pl.:*entreprise*)

La traduction vers le français indique la diversité des rôles sémantiques et fonctions syntaxiques des deux exemples. Dans le premier, le préfixe obéit au pattern archétypique du participe présent : objet direct - verbe - sujet. Dans le second, le préfixe ne saurait assumer cette fonction (ce serait contraire à notre intuition sur le fonctionnement 'normal' de notre monde), mais celle d'un complément circonstanciel ou du point de vue sémantique celle du moyen[16].

Il s'est avéré pour le danois qu'une analyse, mettant en œuvre la partie du discours du préfixe, la construction verbale du verbe dont le participe déverbal est dérivé ainsi que les contraintes de sélection de celui-ci, parvient à désambiguïser la plupart des constructions examinées. Cette analyse évitera du travail et de l'espace dans les dictionnaires, car elle permet de faire l'économie de la lexicalisation de ces participes composés dont le sens est compositionnel, donc prévisible.

Ce ne sont pas seulement les actants (le préfixe et le nom tête du groupe nominal) qui mémorisent des informations nécessaires à l'analyse correcte. D'autres facteurs interviennent, notamment les traits morphosyntaxiques du préfixe et certains traits inhérents au verbe/participe déverbal. Dans les constructions examinées ici, on constate d'une part que ces informations sont neutralisées, et d'autre part qu'elles sont requises en vue d'une traduction satisfaisante.

La neutralisation consiste en ce que les informations ne sont nulle part récupérables dans l'énoncé linguistique, que celui-ci soit écrit ou oral. Il faut donc supposer leur existence à l'état virtuel, soit dans la mémoire des mots, soit dans le savoir partagé des interlocuteurs.

Nous allons aborder dans un premier temps les informations mémorisées dans le verbe, puis celles du préfixe.

Tout particule qui concourt à la construction d'un verbe fini se voit neutralisé dans la construction d'un participe déverbal. Il s'agit des prépositions des verbes transitifs indirects, comme l'exemple évoqué ci-dessus : *forurene* [polluer]. Une variante de cette neutralisation apparaît pour les verbes à particule (non accompagnés de régime). Dans le participe déverbal, la distinction susceptible de modifier le sens d'un verbe préfixé d'un particule et celui d'un verbe que ce même particule suit, est neutralisée. Ainsi, en danois, on distingue *gå på* [én] - *attaquer/agresser qn.* de *pågå - avoir lieu.* Les deux forment le même participe présent déverbal : *pågående.* Ce mot, employé comme épithète, mémorise par conséquent les deux acceptions différentes. Seul l'usage des contraintes de sélection (traits sémantiques) permet de désambiguïser correctement.

Il s'agit aussi, comme en français (cf. Riegel, 1997 : 343 (*illusion évanouie*)), du pronom réfléchi, qui, lui aussi, est neutralisé pendant le processus de préfixation.

Il s'agit, enfin, du mode d'action du verbe. Bien qu'invisible dans le signifiant linguistique, il donne lieu à différentes paraphrases ou différentes traductions. C'est ainsi

16 Le point illustré par ces exemples est peut-être plus clair encore quand il s'agit des noms équivalents : *vandforurening* (pollution de l'eau) et *olieforurening* (pollution par les hydrocarbures).

qu'un participe déverbal dérivé d'un verbe imperfectif correspond à une périphrase en présent :

sygdomstruede landmænd — *des paysans que la maladie menace*
([nom:*maladie-épenthèse* : *s*-part.pa.:*menacer*]-nom pl.:*paysan*)

Un participe déverbal dérivé d'un verbe perfectif correspond, en revanche, à une périphrase en parfait :

sygdomsramte landmænd — *des paysans que la maladie a atteints*
([nom:*maladie-épenthèse* : *s*-part. pa.:*atteindre*]-nom pl.:*paysan*)

À moins de préférer l'équivalent participe qui existe aussi en français [*menacés/atteints (de/par)*], ces exemples montrent que le mode d'action détermine le choix du temps du verbe.

En ce qui concerne les préfixes, ce sont la détermination, le nombre, l'accord et la (forme de la) négation qui sont neutralisées ou modifiées. Là aussi, une traduction appropriée présuppose l'ajout de ces informations. Exemples :

udlandsproducerende virksomheder ([nom:*étranger*-épenthèse :*s*-part.pr.: *produire*]-nom pl.:*entreprise*)	*des entreprises qui produisent* *à l'étranger*[17] Neutralisation de la détermination
arbejdsgiveradministrerede aftaler ([nom:*patron*-part.pa.:*gérer*]-nom pl.: *accord*)	*des conventions gérées par les* *patrons* Neutralisation du nombre et de la détermination
bomberamte tyske byer ([nom:*bombe*-part.pa.:*atteindre*]-adj.: *allemand*-nom pl.:*ville*)	*des villes allemandes que* *des bombes ont atteintes* [*des villes all. bombardées*]
bomberamte japanske byer ([nom:*bombe*-part.pa.:*atteindre*]-adj.: *japonais*-nom pl.:*ville*)	*des villes japonaises que la bombe a* *atteintes*[18] [*des villes jap. bombardées]*

Mentionnons, pour des soucis d'exhaustivité, que les neutralisations accompagnant le processus de préfixation, comprennent aussi l'amalgame de différentes parties du discours (qui se réfèrent au même signifié) ainsi que la neutralisation des contraintes de sélection qui s'appliquent normalement.

17 L'exemple montre qu'il peut y avoir neutralisation de plusieurs traits à la fois.

18 C'est notre savoir de l'histoire contemporaine ainsi que notre connaissance du contexte de l'exemple qui détermineront la bonne traduction. L'essentiel, c'est que le préfixe conserve la même forme, qu'il désigne un signifié singulier ou pluriel.

4. LA TRADUCTION DES PARTICIPES DÉVERBAUX

La conclusion qui s'impose après l'analyse des participes danois composés, c'est qu'une analyse des rapports syntaxiques et sémantiques des participes danois doit nécessairement précéder toute tentative de traduction vers une langue romane comme le français. Sinon, les participes danois, structurellement ambigus et susceptibles de contenir des informations neutralisées, risquent d'être traduits de façon inadéquate.

Dans cette section, nous allons survoler les constructions envisagées afin de présenter les principaux équivalents employés en français pour rendre les participes composés danois. Précisons tout de suite qu'il s'agira d'une simple juxtaposition commentée de quelques exemples. Une véritable analyse pouvant mener à l'établissement de règles de traduction paraît prématurée, vu l'état d'avancement actuel de cette phase du projet.

Nous allons prendre cinq exemples, relevés dans la presse écrite de ces dernières années et censés représenter, sinon de façon exhaustive, du moins une gamme considérable des constructions employées. Les exemples se répartissent en trois groupes, selon la paraphrase choisie :

a) participe déverbal danois vs groupe prépositionnel en français :

det atomdrevne hangarskib CdG (art. def.neut.sg.-[nom:*atome*-part.pa.: *propulser*]-nom sg.:*porte-avions*)	*le (porte-avions) Charles-de-Gaulle, à propulsion nucléaire*

Dans cet exemple le préfixe danois *atom* désigne le moyen de locomotion de ce porte-avions, autrement dit le rôle sémantique. En outre, l'exemple fournit une illustration de cet amalgame de mots auquel il a été fait référence à la fin de la section 3.2. Car ce ne sont pas les atomes qui propulsent ce bâtiment, c'est l'énergie libérée à leur fission qui joue. L'expression correcte et exhaustive serait donc *atomkraftdrevne hangarskib CdG.*

b) participe déverbal danois vs proposition relative en français :

den eneste tilbageværende trussel (art. déf.sg.comm.-adj:*seul*-part.pr.: *rester*-nom sg.:*menace*)	*la seule menace qui continue à exister consiste à...*

C'est sans doute la paraphrase la plus fréquente et celle à laquelle il a déjà été fait allusion ci-dessus. Son emploi présuppose d'un côté l'interprétation correcte des éléments du composé et de l'autre, qu'il n'y ait pas de relatives en cascades. Ce phénomène obère l'identification de l'anaphore au point qu'on préfère souvent d'autres constructions.

c) participe déverbal danois vs groupe adjectival en français :

de værdiskabende kilder (art. déf.pl.-[nom:*valeur*-part.pr.: *créer*]-nom pl.:*source*)	*les sources productrices de valeurs*

de cadmiumforgiftede jordprøver (art. déf.pl.-[nom:*cadmium*-part.pa.: *contaminer*]-nom pl.:*échantillon*)	*les échantillons de sols contaminés au cadmium*
de statsgaranterede lån til SMV (art. déf.pl.-[nom:*État*-épenthèse:*s*-part.pa.:*garantir*]-nom pl.:*prêt*)	*les prêts aux PME garantis par l'État*

Là où le rapport entre le participe déverbal et son préfixe est neutralisé en danois, il est partiellement explicite en français par le moyen d'une préposition. Cette clarté est pourtant toute relative, car la préposition *par*, pour prendre un exemple, peut désigner soit l'agent, comme c'est le cas de l'exemple cité, soit le moyen. L'analyse qu'une paraphrase relative fournirait demeure donc potentiellement nécessaire.

Mentionnons une autre stratégie qui consiste à ne pas traduire le préfixe. Nous ne disposons pas d'exemples de l'emploi de cette stratégie pour la traduction entre le français et le danois. Il n'y aurait sans doute pas de difficultés à en repérer. Un exemple venant d'une traduction du danois vers l'anglais fournit cependant une illustration de cette stratégie. On sait que l'action du verbe *udhule* [creuser] demande un agent (humain), un patient (un tronc) et enfin un moyen (un outil). Le nombre d'actants préfixés étant limités à un seul, on a pour le participe passé le choix entre le sujet (l'agent) ou le moyen. L'un ou l'autre est sous-entendu. Dans l'exemple suivant, le danois avait préféré préfixer l'outil, alors qu'en anglais celui-ci n'était pas exprimé :

en økseudhulet lindestamme	*a hollowed-out lime trunk* [un tronc de tilleul creusé]

5. (EN GUISE DE) CONCLUSION

L'exposé qu'on vient de lire a voulu décrire une série de contraintes auxquelles est sujet l'usage des participes français et danois. Pour les participes simples, il s'agit notamment de contraintes de formation et de combinatoire.

Il semble que ces contraintes s'expliquent le mieux en faisant référence à un savoir implicite ou bien à un embryon de mémoire qui conserve les propriétés issues de la construction des verbes dont les participes sont dérivés.

Il est à noter que les deux langues sont sujettes à des contraintes qui sont souvent identiques.

Puisque la formation lexicale danoise autorise une préfixation inconnue en français, la traduction entre ces deux langues et a fortiori entre les familles dont ces deux langues sont issues, doit tenir compte d'une série de problèmes qui, eux aussi, font appel à la mémoire des mots. Il s'agit de deux groupes de problèmes :

- ambiguïté structurelle, qu'il s'agisse de la mémoire de la construction syntaxique ou de la mémoire des contraintes sémantiques;

récupération du savoir implicite contenu dans le participe déverbal ou dans son préfixe.

Une traduction entre ces deux langues exige que ces problèmes soient résolus de manière satisfaisante.

6. RÉFÉRENCES

ARRIVÉ, Michel, GADET, Françoise et Michel GALMICHE (1986) : *La grammaire d'aujourd'hui – guide alphabétique de linguistique française*, Paris, Flammarion, 720 p.

GREVISSE, Maurice et André GOOSSE (1993) : *Le bon usage*, Paris/Louvain-la-Neuve, DeBoeck et Duculot, 1762 p.

HALMØY, Odile (1984) : «À propos de l'adjectif en -ant, dit "verbal"», *Revue Romane*, 19.1, Copenhague, pp. 48-64.

KJÆRSGAARD, Poul Søren (1995) : *Fransk grammatik - i hovedtræk* [Grammaire française], Odense, Odense Universitetsforlag, 310 p.

KJÆRSGAARD, Poul Søren (1995) : «Disambiguering af enkle og sammensatte participialer i dansk» [La désambiguïsation des participes déverbaux simples et composés en danois], Poul Søren Kjærsgaard et Lene Schøsler (Eds), *UDOG – Udforskning af dansk ordforråd og grammatik* [Recherches sur le vocabulaire et la grammaire danois], coll. Udog, n° 3, Odense, pp. 9-22.

KJÆRSGAARD, Poul Søren (1996): «Danske participialer og valens» [Les participes danois et la valence], Karin Van Durme (Ed.), *Adjektivernes valens* [La valence des adjectifs], coll. Odense Working Papers in Language and Communication, n° 12, Odense, pp. 49-84.

KJÆRSGAARD, Poul Søren et Bjarne LE FEVRE JAKOBSEN (1997, à paraître) : «Sondringen mellem participiale deverbaler og adjektiver» [La distinction entre participes déverbaux et adjectifs en danois], Bente Maegaard et Bolette Sandford Pedersen (Eds), *UDOG – Udforskning af dansk ordforråd og grammatik* [Recherches sur le vocabulaire et la grammaire danois], coll. Udog, n° 6, Copenhague.

PERLMUTTER, David M. and Carol G. ROSEN (Eds) (1983-90) : *Studies in Relational Grammar*, I-III, The University of Chicago, Chicago, Ill.

PÜMPEL-MADER, Maria, GASSNER-KOCH, Elsbeth und Hans WELLMANN (1992) : *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache*, Tome 5, Adjektivkomposita und Partizipialbildungen, (Komposita und kompositionsähnliche Strukturen), Berlin/New York, Walter de Gruyter, 340 p.

RIEGEL, Michel, PELLET, Jean-Christophe et René RIOUL (1997) : *Grammaire méthodique du français*, Paris, Presses Universitaires de France, 646 p.

Remerciements

L'auteur tient à remercier ses collaborateurs de ce projet pour leur aide et leurs commentaires durant l'élaboration de cet article, Bjarne le Fevre Jakobsen, Jens Ahlmann Hansen et Peter Stewart.

Les exemples français fournis sont, dans la mesure du possible, vérifiés à l'aide de la base textuelle de l'INaLF, Frantext.

EXPLORATION DE CLASSIFIEURS CONNEXIONNISTES POUR L'ANALYSE DE TEXTES ASSISTÉE PAR ORDINATEUR

Jean-Guy MEUNIER, Ismaïl BISKRI, Georges NAULT, Moses NYONGWA

LANCI, Université du Québec à Montréal, Montréal, Canada

1. PRÉSENTATION GÉNÉRALE

Cet article présente un volet de la recherche et du développement effectués dans le cadre du projet Franco-Québécois intitulé «Les classifieurs émergentistes et le traitement de l'information». L'objectif général de ce projet est d'explorer les approches classificatoires statistico numériques pour l'analyse de l'information se présentant en langue naturelle, texte, fiche, documents etc. (leur pertinence également). Dans le présent travail, les classifieurs explorés ont été les réseaux de neurones, les analyses multidimensionnelles, les algorithmes génétiques, les champs de Markov. Concrètement, le but visé par ces recherches est de trouver une méthode efficace et économique d'appariement de tels types de données informationnelles. Dans le cas du texte, il s'agit de formation de classes de segments textuels qui se ressemblent en raison d'un critère particulier choisi dans l'expérimentation. Ces classes peuvent alors être soumises à d'autres analyseurs qui eux ouvrent la voie à des fonctions d'analyses plus complexes. Celles-ci deviennent alors des moteurs de fouille du type hypertextualisation automatique, extraction de connaissances, indexation, analyse terminologique etc. Dans le cas de fiches descriptives (genre : dossiers de patients), il s'agit aussi de formation de classes de fiches présentant des similarités, par exemple entre patients ou entre symptômes. Ces classes sont ensuite soumises à des analyses plus fines et intégrées dans des systèmes d'assistance dynamique aux diagnostics.

Un des résultats périphériques mais des plus prometteurs de cette collaboration est l'entente explicite d'un certain nombre de chercheurs pour collaborer à un protocole ou une plate-forme commune afin de permettre aux divers logiciels francophones utilisés dans ce projet d'entrer en interaction. Trop de logiciels français restent sur les tablettes faute de plate-forme pour les mettre en interaction. Le LANCI propose une plate-forme, ALADIN (Seffah et Meunier, 1995), qui constitue une première tentative, prototypale, pour mettre concrètement en interaction ces logiciels. Les laboratoires LANCI et TIMC-IMAG travaillent dans le cadre du présent projet à la réalisation de cet objectif. Dans cet article, nous présentons brièvement l'exploration et l'expérimentation d'une approche neuronale pour l'analyse terminologique dans les grand corpus.

2. COLLABORATION FRANCO-QUÉBÉCOISE

Ce projet a été réalise dans le cadre d'un programme Franco-Québécois, sous la co-tutelle du ministère de l'Enseignement supérieur et de la Recherche pour la partie française (M.E.N.E.S.R.I.P., Délégation à l'Information Scientifique et Technique, Programme Ingénierie linguistique et de la connaissance) et du ministère de la Recherche et du ministère des Affaires Internationales, de l'Immigration et des communautés Culturelles pour la partie Québécoise.

Du côté français, ce projet s'est déployé en une collaboration portant sur l'assistance au diagnostique médical dynamique où participent Dr. Danel, du CHU de Grenoble, V. Rialle (et son équipe) du TIMC IMAG de Grenoble.

Du côté Québécois, en raison de la dimension textuelle, ce projet a permis de dynamiser grandement l'intégration de l'équipe à plusieurs projets AUPELF-UREF FRANCIL. sur les systèmes logiciels d'assistance à la terminologie (ARC A3). Ce qui s'est à son tour déployé en une collaboration intense et suivie avec l'université de Paris-Sorbonne (Laboratoire CAMS-LALIC : dir. J.P. Desclés) ainsi que l'IDIST de Lille 3.

3. CLASSIFIEURS NEURONAUX ET ASSISTANCE À LA TERMINOLOGIE

De nos jours, un nombre croissant d'institutions accumulent très rapidement des quantités de documents qui ne sont souvent classés ou catégorisés que très sommairement. Très vite, les tâches de dépistage, d'exploration et de récupération de l'information présente dans ces textes, c'est-à-dire des «connaissances», deviennent extrêmement ardues, sinon impossibles. Pour y faire face, il devient nécessaire d'explorer de nouvelles approches d'aide à la lecture et à l'analyse de textes assistées par ordinateur (LATAO).

Extraction et classification

La littérature technique relative au traitement de l'information textuelle a montré qu'il était possible d'explorer des outils d'extraction des connaissances dans des textes (*data mining*). Pour les chercheurs dans le domaine de LATAO, cette problématique n'est pas nouvelle. Dans la recherche antérieure, plusieurs techniques et méthodes ont déjà été proposées pour tenter d'organiser le contenu d'un texte en des configurations interprétables. Ces méthodes, souvent moins fines certes que les approches linguistiques et conceptuelles n'en permettent pas moins un premier parcours général et robuste du texte. Elles sont en mesure, par exemple, d'identifier dans un corpus des classes ou des groupes de lexèmes qui entretiennent entre eux des associations dites de cooccurrence et donc de détecter leurs réseaux sémantiques. Et les recherches actuelles commencent d'ailleurs à les privilégier de plus en plus. Parmi les modèles les plus couramment utilisés, on trouve habituellement l'analyse des cooccurrences, l'analyse corrélationnelle, l'analyse en composante principale, l'analyse en groupe, l'analyse factorielle, l'analyse discriminante, etc. Malgré le succès qu'elles ont obtenu, on a dû constater que ces méthodes particulières posent deux problèmes importants.

Premièrement, les modèles classiques ne peuvent traiter que des corpus stables. Toute modification du corpus exige une reprise de l'analyse numérique. Ceci devient un problème majeur dans des situations où le corpus est en constante modification (par exemple, les reposoirs de l'autoroute électronique). Deuxièmement, les types de résultats qu'ils produisent ne sont pas sans problèmes théoriques. Ils posent des problèmes d'interprétation linguistique importants (Church et Hanks, 1990). Les associations de mots dans les classes ne sont pas toujours facilement interprétables. Pourtant, malgré leurs limites, ces approches ont été reconnues des plus utiles pour l'extraction des connaissances et plus particulièrement les connaissances terminologiques. D'une part, ces stratégies classificatoires permettent une immense économie de temps dans le parcours exploratoire d'un corpus, et à ce titre, elles sont incontournables lorsqu'on est confronté à de vastes corpus textuels. D'autre part, elles servent d'indices pour détecter rapidement certains liens sémantiques et textuels. Cependant, lorsqu'associées à des stratégies linguistiques plus fines et intégrées dans des systèmes hybrides (i.e., avec analyseurs linguistiques d'appoint), elles livrent une assistance précieuse pour des analyses globales. Elles permettent un premier déblaiement général du texte. Peuvent alors suivre des analyses plus fines.

Les recherches récentes permettent de penser qu'on peut améliorer ces techniques de classification de l'information. En effet, de nouveaux modèles classifieurs dits émergentistes commencent à être explorés pour ce type de tâche. Ils ont pour fondement théorique que le traitement «intelligent» de l'information est avant tout associatif et surtout adaptatif. Parmi ces modèles dits «de computation émergente» on distingue les modèles «génétiques» , markoviens et surtout connexionnistes. Parmi ces derniers, on trouve une grande variété de modèles, entre autres, les modèles matriciels linéaires et non linéaires, les modèles thermodynamiques, et les modèles basés tantôt sur la compétition, tantôt sur la rétropropagation, mais surtout sur des règles complexes d'activation et d'apprentissage. Les principaux avantages de ces modèles tiennent au fait que leur structure parallèle leur permet de satisfaire un ensemble de contraintes qui peuvent être faibles et même, dans certains cas, contradictoires et de généraliser leur comportement à des situations nouvelles (le filtrage), de détecter des régularités et ce, même en présence de bruit. Outre les propriétés de généralisation et de robustesse, la possibilité pour ces modèles de répondre par un état stable à un ensemble d'inputs variables repose sur une capacité interne de classification de l'information.

Cependant, tous ces modèles classifieurs émergentistes opèrent souvent sur des données bien contrôlées et qui toutes doivent être présentes au début et tout au long du traitement. De plus, ils exigent souvent divers paramètres d'ajustement qui relèvent souvent d'une description statistique du domaine. Il s'en suit que les résultats de classification obtenus sont valides pour autant qu'ils portent sur les données bien contrôlées où peu de modifications sont possibles. Si, après la période d'apprentissage, pour quelque raison que ce soit, les systèmes sont confrontés à des données qui n'étaient pas prévues dans les données de départ, ils auront tendance à les classer dans les prototypes déjà construits, donc à produire une sous-classification.

Or, dans le domaine du texte, nous sommes confrontés a des corpus en constante modification. Chaque nouvelle page peut possiblement contenir des informations que le système peut ne jamais avoir rencontrées, et donc qu'il ne peut se permettre de classer dans ses prototypes antérieurement construits. Il faut donc, outre la dynamicité (incrémentialité) de l'apprentissage, un système qui soit aussi plastique, c'est-à-dire qui s'adapte à de

nouvelles données. Et on voit apparaître depuis quelque temps des recherches qui sont de plus en plus sensibles à cette dimension (Burr, 1987; Veronis et al., 1990; Balpe et al., 1996, etc.). Et c'est dans cette perspective que la présente recherche a été effectuée. Nous en présentons ici les résultats préliminaires à l'occasion d'une application d'analyse terminologique.

4. LA MÉTHODE

Dans sa forme concrète et expérimentale, la recherche a consisté à explorer un modèle connexionniste pour extraire de l'information de type terminologique sur des fichiers textuels. La réalisation de cette recherche comporte les étapes suivantes :

1- une modélisation d'un traitement connexionniste du texte;

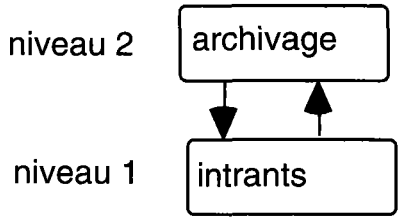
2- une expérimentation d'une chaîne de traitement sur des textes.

4.1 Modélisation d'un traitement connexionniste de textes

Il n'existe pas, à notre connaissance, un grand nombre de modèles connexionnistes pouvant confronter simultanément les problèmes de dynamicité et de plasticité de l'information. La dynamicité est la capacité du système à traiter de manière adaptative les informations qu'il reçoit. Quant à la plasticité, il s'agit de la capacité du système à traiter de l'information pour laquelle il n'avait pas été paramétrisé. Un des modèles qui avait ces visées est ART 1 (Grossberg et Carpenter, 1987). En effet, dans sa définition originale le modèle ART 1 se veut un système classifieur auto-organisationnel, sans supervision, pouvant opérer sur des stimulis non contrôlés et bruités.

Ce modèle a évolué à travers les années. Il a donné naissance à de nombreuses variantes de la règle de transmission et de la règle d'apprentissage. Il a introduit des facteurs de multiplication dans l'activation et des facteurs de dégradation dans l'encodage de l'information. L'une de ses prétentions importantes est qu'il est en mesure de traiter de manière adaptative des stimulis qui sont changeants (plasticité), c'est-à-dire qui ne font pas partie d'un corpus contrôlé d'avance. L'objectif ultime de ce modèle est de créer une grande stabilisation dans la représentation des patrons de stimulis. Au début, ART 1 ne pouvait traiter que des informations de nature binaires. Plus tard, le modèle ART 2 accepte des informations dont les valeurs ne sont plus discrètes ou binaires. Enfin dans ART 3 le modèle a consolidé ses stratégies et offre un traitement plus fiable.

L'idée centrale du modèle ART est celle d'un système d'interaction entre deux niveaux qui entrent en résonance mutuelle.



Le système reçoit dans un premier niveau 1 des stimulis qui sont envoyés, mais aussi modifiés (selon une distribution et un poids particulier) au deuxième niveau 2, qui est un niveau d'archivage.

Arrive donc au deuxième niveau un pattern différent de ce qui était à l'intrant. Il y a alors comparaison par un processus dit de résonance. Si le nouveau pattern n'a aucune ressemblance avec les anciens, il sera alors conservé et servira de gabarit ou de prototype avec lequel les intrants nouveaux seront éventuellement comparés. En fait, le pattern au niveau 2 servira de modèle de comparaison avec les nouveaux intrants. S'il diffère de ces derniers, un autre pattern sera essayé jusqu'à ce qu'une correspondance soit acceptable (selon certains paramètres) et s'il y a correspondance acceptable l'intrant sera alors classé avec le prototype. Mais s'il n'est pas acceptable, le nouveau pattern sera considéré comme un prototype en émergence et il servira éventuellement de nouveau gabarit aux autres intrants que le système rencontrera.

La correspondance entre le pattern prototypal et le pattern intrant est la résonance. Au fur et à mesure que l'apprentissage se poursuit, il y a consolidation de cette résonance. L'adaptabilité survient par la modification constante des interconnexions entre les niveaux. Pour réaliser cette interaction le système doit être contrôlé par divers paramètres qui assurent la solidité du traitement.

4.2 L'application du modèle ART

La deuxième étape de la recherche est une expérimentation de ce modèle sur des textes et sur une tâche spécifique. Un texte regorge de connaissances de plusieurs types qui peuvent en être extraites. On ne lit pas un texte uniquement pour savoir qui a fait quoi (connaissances du monde). On le lit pour connaître des faits, certes, mais aussi des actions, des valeurs, des jugements etc. (Meunier, 1996). Une des connaissances qu'on cherche est aussi de nature métalinguistique, i.e. qui porte sur la nature de la langue même du texte, par exemple, son style, sa terminologie, son lexique, etc. Dans la présente recherche, nous avons choisi d'explorer l'extraction des connaissances métalinguistiques de types terminologique et sémantique. Pour ce faire nous avons développé une *chaîne de traitement* modulaire qui tente d'intégrer les étapes du travail analytique du terminologue ou du documentaliste devant travailler à l'identification des champs lexicaux d'un terme particulier. Par exemple, la chaîne de traitement pourrait aider le terminologue à identifier rapidement que, dans un texte, le mot CODE peut avoir des champs sémantiques différents où ce mot prend le sens de CODE civil, CODE de la route, CODE informatique, CODE de la construction, etc.

Cette chaîne de traitement a été appliquée à deux textes spécifiques qui nous servent ici d'illustration : le premier de 900 pages, la *Convention de la Baie James* d'Hydro Québec; le second, la revue *Spirale* (Belgique) de quelque 180 pages.

L'expérimentation sur les textes

Dans la première étape de sa gestion, le texte est reçu et traité par des modules d'analyse de la plate-forme ALADIN-TEXTE. Cette plate-forme est un atelier qui utilise des modules spécialisés dans l'analyse d'un texte. Dans un premier temps, un filtrage sur le lexique du texte est fait. Par divers critères de discrimination, on élimine du texte certains mots accessoires (mots fonctionnels ou statistiquement insignifiants, etc.) ou ceux qui ne sont pas porteurs de sens d'un point de vue strictement sémantique, et dont la présence pourrait nuire au processus de catégorisation, soit parce qu'ils alourdiraient indûment la représentation matricielle, soit parce que leur présence nuirait au processus interprétatif qui suit la tâche de catégorisation. Vient ensuite une description morphologique minimale de type lemmatisation. Cette opération consiste à remplacer chaque mot par son équivalent canonique. (e.g. aimerions --> AIMER). Ce processus se justifie par le fait que les déclinaisons propres à la grammaire ou à la syntaxe d'une langue n'affectent en rien le contenu sémantique réel des termes. De la même façon, remplacer un mot décliné (soit dans sa forme verbale, adverbiale, adjectivale, pronominale ou autres) par sa forme nominale n'a aucun impact significatif sur le contenu sémantique principal de ce dernier. Ces dimensions morphologiques touchent surtout des modalités tels : le genre, l'aspect, le temps, etc.

Puis une transformation est opérée pour obtenir une représentation matricielle du texte. Cette transformation est encore effectuée par des modules d'ALADIN explicitement dédiés à cette fin. On produit ainsi un fichier indiquant pour tout lemme choisi sa fréquence dans chaque segment du texte. Suit ensuite un post-traitement pour construire une matrice dans un format acceptable par les réseaux de neurones. Dans la présente expérimentation, la précédente matrice est alors soumise aux classifieurs ART. Selon le réseau utilisé (ART 1 ou FUZZY-ART), la matrice générée peut être constituée exclusivement de données binaires (ART 1) ou de données non binaires (FUZZY-ART). Dans le cas du réseau ART 1, les données subissent alors une réduction, puisque la fréquence d'apparition des lemmes est alors remplacée par une simple indication de présence ou d'absence.

Les réseaux de type ART se prêtent particulièrement à ces contraintes. Le Fuzzy-Art semble donner des résultats supérieurs car le processus de catégorisation produit moins de classes contenant un seul élément caractéristique (un seul lemme), ce qui du point de vue du terminologue n'est d'aucune utilité.

Les résultats de la classification neuronale

Appliqué à la matrice précédente, ART génère des classes de SEGMENTS qui présentent entre eux une certaine similarité lexicale. Autrement dit, chaque classe de segments constitue pour ART un «prototype» qui est donc caractérisée par les termes qui sont présents également dans tous les segments du texte. Une fois trouvées ces classes de segments présenant une similarité lexicale, on en extrait, pour chacune, le lexique, c'est -à-dire on trouve pour les termes qui la caractérise.

Ensuite, on choisit un terme particulier, et on étudie les classes dans lesquelles il apparaît. C'est ainsi, par exemple, que dans l'analyse sur la revue *Spirale*, nous avons obtenu la classification suivante pour les termes *rapport* et *tête*

Le terme *rapport*
apparaît dans les classes 28 35, 39 40 54,
la classe 28 : les segments 71-73
contient : *choix connaissance, document, façon fiction personnage, rapport, savoir et travail*
la classe 35, les segments 89-92 qui contiennent les mots
autres connaissances doute formes image processus production et rapport
la classe 39 : 100 101 qui contiennent les mots
autres élèves enseignant ensemble genre, jeunesse rapport roman
la classe 40 : 102 103 qui contiennent les mots
écrit, écriture élémentaires, jeunesse, monde problème rapport réel, scolaire, situation
la classe 54 : 59 64 qui contiennent les mots
auteurs, autres, discours, jeunesse, lecture, mode, question, rapport, rôle, temps, vie

Sur la distribution de ce terme dans les classes de segments trouvées, on peut tenter l'interprétation sémantique suivante. On peut dire que le terme *rapport* est utilisé dans deux ensembles de contextes relativement différents.

Un premier pointe vers le concept de *rapport* comme *document* où est déposé de l'information (classe 28 dans les segments 71-73).

Un deuxième pointe vers le concept des *liens entre des individus et autres chose* (39, 40, 54).

Enfin la classe 35 n'est pas clairement situable dans l'un ou l'autre sens précédent.

Cependant, on voit bien que dans ce texte, ces deux significations sont les deux seules possibles. Pour un terminologue, le terme rapport dans la revue *Spirale* # 15 n'est pas employé dans le sens suivants : d'une proportion, c'est-à-dire d'un rapport logique, d'un rapport financier, d'une maison de rapport, d'une communication ou encore d'une perspective, etc.

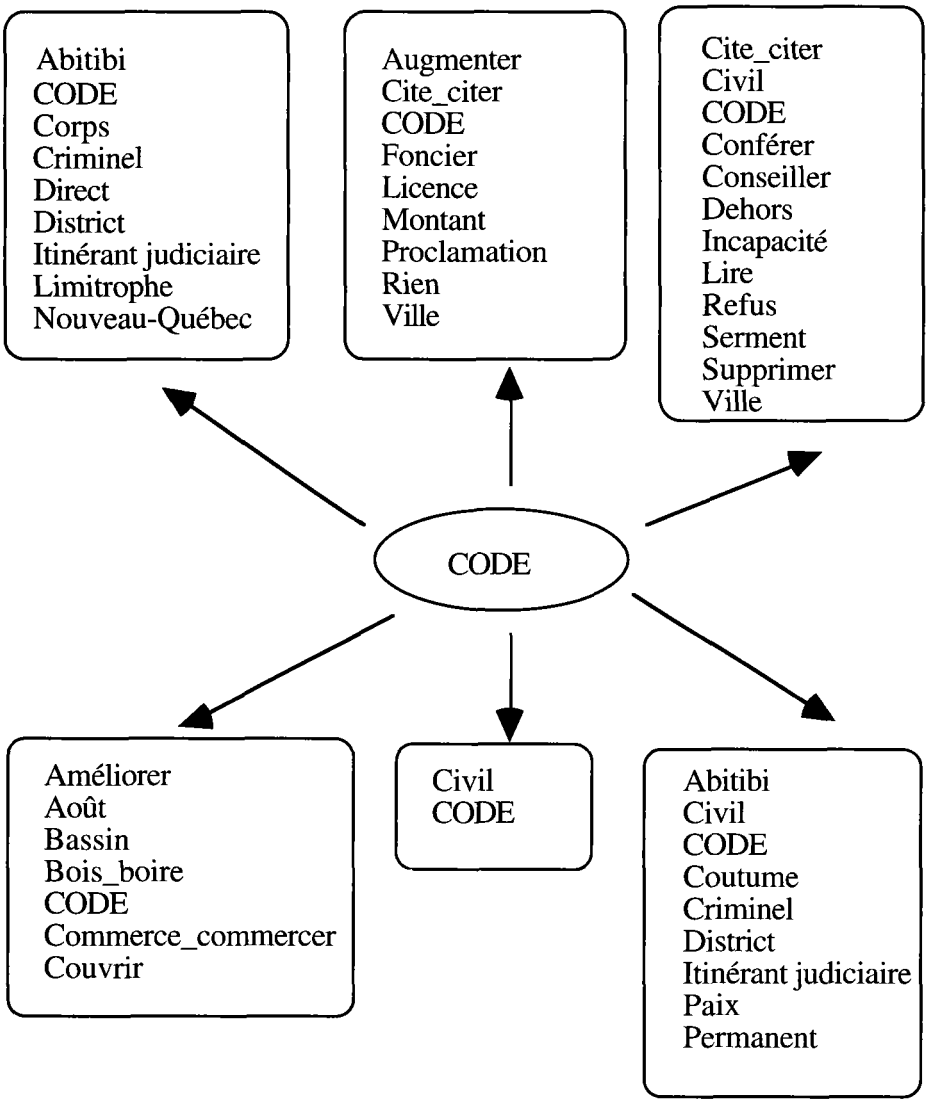
Une analyse similaire peut être tentée sur le terme *tête*.

Nous retrouvons le terme *tête* dans les classes dans la classe 20 et la classe 58.
La classe 20 : les segments 53 et 107.
façon, moment, place, rôle, savoir, tête.
La classe 58 : le segment 93
aide, bon, comprendre, connaissance, École, ensemble, histoires, jour, lecteur, loisirs, narratif, père, perspective, schéma, situation, souris, suite, tête, textuels, titre, traitements.

Ces deux classes montrent que le terme *tête* est utilisé dans deux sens différents. Pour la classe 20 le sens du terme tête est proche de la signification «leader» or dans la classe 58 la signification est plus proche de l'interprétation «intelligence».

Le *post-traitement*

La présentation linéaire de ces résultats rend l'interprétation difficile à réaliser. On peut imaginer une meilleure convivialité de la configuration des résultats. En effet, une légère transformation peur leur donner une lecture plus facile. À titre d'exemple nous présentons ci-après la configuration d'un terme plus riche d'acception sémantique que nous avons trouvé dans le texte d'Hydro Québec : la *Convention de la Baie James* (870 p.). Dans ce texte, le mot CODE se retrouve dans plusieurs classes. Il est alors possible pour ce lexème de dessiner le graphe des classes dans lesquelles il apparaît.



On voit ainsi apparaître la différence des réseaux sémantiques de ce terme, qui tantôt est utilisé dans le sens de code civil, tantôt dans le sens de code de comportement, tantôt dans le sens du code criminel, etc. Pour un terminologue, ce graphe est intéressant. Il sert d'indice au réseau sémantique de ce terme, et donc de ses acceptions dans le discours particulier. La thèse associative classique explique ce fait en postulant que si deux termes se retrouvent ensemble dans un même contexte, c'est que leurs contenus sémantiques ou conceptuels sont associés.

5. INTERPRÉTATION ET VALIDATION DES RÉSULTATS

Les résultats produits sont alors prêts à être validés. Une méthode de validation inter-juges a été utilisée pour évaluer la qualité et la pertinence de la classification opérée par le réseau de neurones. La méthode consiste essentiellement à comparer les évaluations faites et les appréciations effectuées par les différentes personnes impliquées dans l'évaluation des résultats. Le travail a consisté essentiellement en une analyse des classes produites pour déterminer leurs pertinences d'un point de vue terminologique. De façon générale, l'analyse a montré que plus de 80 % des classes constituées seraient pertinentes et utilisables dans le processus de création des fiches terminologiques. Mais un travail de validation sur un corpus banc d'essai est maintenant nécessaire.

6. CONCLUSION

L'objectif de notre recherche était de permettre l'extraction de connaissances terminologiques à partir du texte plein. Cette extraction devait, de plus, pouvoir se faire sur un corpus en évolution constante (plasticité) et les catégorisations effectuées se devaient de rester pertinentes et utilisables. Le processus de catégorisation quel qu'il soit se devait d'opérer sans supervision aucune (adaptabilité) et sans faire appel à des connaissances préformées ou prédigérées, celles-ci n'étant simplement pas disponibles dans le cas qui nous concerne.

Une méthode connexionniste comme solution au problème de l'extraction terminologique sur des textes entiers a été expérimentée avec des résultats très encourageants. L'approche choisie montre un intérêt certain et des avantages indéniables. Par exemple, le gain de temps estimé par rapport au travail manuel requis pour effectuer un travail terminologique équivalent est considérable. De plus, la précision et la richesse des suggestions faites par le système ne sont en aucune mesure comparables avec ce qu'on obtient avec les méthodes actuelles.

L'approche choisie s'inscrit parfaitement dans l'optique d'une solution opérationnelle aux problèmes flagrants et réels qui minent l'industrie de la langue et toute industrie qui implique la manipulation et la classification de masses importantes de documents.

Plusieurs variantes de l'expérimentation sont possibles et envisagées. On pourra opérer les classes autrement que par l'intersection entre les segments. Par exemple, tirer des informations pertinentes de l'union des unifs présents dans les segments regroupés. On pourrait sûrement obtenir des résultats encore plus probants en tenant compte de la dépendance entre les variables (ou unifs) pris en considération. Un pré-processing sémantique (non approfondie pour conserver l'avantage de temps) pourrait amener des améliorations considérables. Plusieurs variantes des filtres de pré-traitement utilisés sont à l'étude. Le filtrage à la sortie des classes ne représentant pas une richesse d'information suffisante (selon un critère donné, ex. : pas assez d'unifs dans cette classe) améliorerait les résultats.

Plusieurs problèmes restent à résoudre (grandeurs fixes des intrants, dégradation du temps d'apprentissage avec le nombre des intrants, interprétation, etc.).

Un module d'interprétation des résultats (avec interface graphique) est en cours de développement.

RÉFÉRENCES

BALPE, J. P., LELU, A., PAPY, F., & I, S. (1996) : *Techniques avancées pour l'hypertexte,* Paris, Hermes.

BURR, D. J. (1987) : «Experiments with a Connectionnist Text Reader», *IEEE First International Conference on Neural Networks*, San Diego, pp. 717-724.

CARPENTER, G. & G. GROSSBERG (1991) : «An Adaptive Resonnance Algorithm for Rapid Category Learning and Recognition», *Neural Networks*, 4, pp. 493-504.

CHURCH, K., GALE, W., HANKS, P. & D. HINDLE (1989) : «Word Associations and Typical Predicate-argument Relations», *International Workshop on Parsing Technologies,* Carnegie Mellon University, Aug. 28-31.

CHURCH, K. W. & P. HANKS (1990) : «Word Asociation Norms, Mutual Information, and Lexicography», *Computational Linguistics*, 16, pp. 22-29.

DELISLE, S. (1994) : *Text Processing Without A Priori Domain Knowledge : Semi Automatic Linguistic Analysis for Incremantal Knowledge Acquisition*, PhD Thesis, Ottawa University.

GARNHAM, A. (1981) : «Mental Models and Representation of Texts», *Memory and Cognition*, 9, pp. 560-565.

GREFENSTETTE, G. (1992) : «Sextant : Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis», *Proc. of the 30th Annual Meeting of the ACL*, pp. 324-326.

GREFENSTETTE, G. (1992) : «Use of Syntactic Context to Produce Term Association Lists for Text Retrieval», *Proc. of SIGIR 92 ACM*, Copenhagen, June 21-24.

GROSSBERG, S. & S. CARPENTER (1987) : «Self Organization of Stable Category Recognition Codes for Analog Input Patterns», *Applied Optics*, 26, pp. 4919-4930.

JACOBS, P. & U. ZERNIK (1988) : «Acquiring Lexical Knowledge from Text A Case Study», *Proceedings of AAA1 88*, St Paul, Min.

KOHONEN, T. (1982) : «Clustering, taxonomy and topological Maps of Patterns», *IEEE Sixth International Conference on Pattern Recognition*, pp. 114-122.

LEBART, L. & A. SALEM (1988) : *Analyse statistique des données textuelles*, Paris, Dunod.

LELU, A. (1995) : «Hypertextes : la voie de l'analyse des données», L. Bolasco, S. L. A. Salem (Eds), *Anilisi statistica dei dati testuali* , vol. 2, Rome, CISU, pp. 85-96.

LIN, X., SOERGEL, D. & G. MARCHIONINI (1991) : «A Self Organizing Semantic Map for Information Retrieval», *SIGIR 91*, Chicago, Illinois.

MEUNIER, J.-G. (1996) : «Théorie cognitive : son impact sur le traitement de l'information textuelle», V. Riale et D. Fisette, *Penser l'esprit des sciences de la cognition a une philosophie cognitive*, Presses de Université de Grenoble, pp. 289-305.

MOULIN, B. & D. ROUSSEAU (1990) : «Un outil pour l'acquisition des connaissances à partir de textes prescriptifs», *ICO*, Québec 3 (2), pp. 108-120.

REGOCZEI, S. et al. (1988) : «Creating the Domain of Discourse : Ontology and Inventory». Gaines &. Boose (Eds), *Knowledge Acquisition Tools for Experts and Novices*, Academic Press.

REGOCZEI, S. & G. HIRST (1989) : On extracting knowledge from Text. Modeling the Architecture of Language Users. (TR CSRI 225), Computer Systems Research Institute, University of Toronto.

SALTON, G. (1988) : «On the Use of Spreading Activation», *Communications of the ACM*, vol 31 (2).

SALTON, G., ALLAN, J. & C. BUCKLEY (1994) : «Automatic Stucturing and Retrieval of Large Text File», *Communications of the ACM*, 37 (2), pp. 97-107.

SEFFAH, A. & J.-G. MEUNIER (1995) : «ALADIN : un atelier orienté objet pour l'analyse et la lecture de textes assistée par ordinateur», *International Conference on Statistics and Texts,* Rome.

TAPIERO, I. (1993) : *Traitement cognitif du texte narratif et expositif et connexionnisme : expérimentations et simulations*, Université de Paris VIII.

THRANE, T. (1992) : «Dynamic Text Comprehension», J. O. S. Jansen, H. Prebensen, T. Thrane (Eds), Copenhaguen, Museum Tuscalanum Press.

VERONNIS, J., IDE, N. M. & S. HARIE (1990) : «Utilisation de grands réseaux de neurones comme modèles de représentations sémantiques», *Neuronimes*.

VIRBEL, J. (1987) : «L'apport de connaisances linguistiques à l'interprétation des structures textuelles», *Structure des documents, Bigre++Globule,* 53, pp. 77-97.

VIRBEL, J. E., F. PASCUAL (1992) : *La lecture assistée par ordinateur, Raport de recherche,* Toulouse, Laboratoire IRIT.

VIRBEL, J. (1993) : «Reading and Managing Texts on the Bibliothèque de France Stations», P. Delany & P. Landow (Eds), *The Digital Word : Text Based Computing in the Humanities, Cambridge*, Mass., MIT Press.

YOUNG, T. & T. CALVERT (1987) : *Classification, Estimation, and Pattern Recognition*, Amsterdam, Elsivier.

ZARRI, G. P. (1990). «Représentation des connaissances pour effectuer des traitements inférentiels complexes sur des documents en langage naturel», Office de la langue française (Ed.), *Les industries de la langue. Perspectives 1990*, Gouvernement du Québec.