PARTIE I

Traduction automatique et traduction assistée

Présidents : Marcel Voisin
Christian Delcourt
Roger Goffin

1

Quelle automatisation de la traduction peut-on souhaiter et réaliser sur les stations de travail individuelles?

Christian BOITET

GETA, Institut IMAG (UJF et CNRS), Grenoble, France

Résumé

Une brève revue de la situation actuelle permet de dégager une classification des types de traduction (re-création, localisation, traduction-diffusion, traduction-dépistage) et de déterminer dans quelle mesure on peut automatiser la « fonction traduisante » proprement dite, c'est-à-dire la production automatique de traductions « brutes » ou « grossières » (destinées à la diffusion après révision ou au dépistage). Comme les possibilités et les limites de la TAO « classique » (TA) sont souvent méconnues, nous indiquerons au passage quelques règles de « bon usage » de la TA.

En ce qui concerne les « stations de travail » associées, l'accès à des traductions grossières (dépistage) doit pouvoir être réalisé sur des matériels très variés, allant du terminal Minitel à un poste dédié. Par contre, la révision de traductions brutes nous semble demander le même type d'environnement que la traduction professionnelle industrielle (en groupe).

Quand on ne peut raisonnablement proposer une technique de TA, on peut souvent fournir des aides au traducteur humain. Dans un cadre professionnel, il faut prévoir un environnement pour le travail en petite équipe (*groupware*), l'accès aux traductions passées, l'intégration dans un système de documentation informatisée, et de puissantes bases de données lexicales et terminologiques. Les stations de travail et les logiciels utilisés peuvent être de haut de gamme, et/ou dédiés.

S'il s'agit de traduction occasionnelle, il faut des environnements moins ambitieux, mais utilisables en conjonction avec une grande variété d'outils du commerce (texteurs, documenteurs, tableurs, grapheurs, SGBD...). Du point de vue informatique, la conception et la réalisation de ce type d'outils est particulièrement intéressante, à cause des exigences de généricité, de portabilité et de simplicité. Bien sûr, il faut viser des matériels accessibles au grand public.

À côté des stations pour le veilleur, pour le réviseur-traducteur professionnel, et pour le traducteur occasionnel, on peut enfin imaginer (mais sera-ce pour l'an 2001?) des stations pour rédacteurs unilingues ne pouvant utiliser les services d'un traducteur, désirant cependant être traduits dans une ou plusieurs langues, et acceptant pour cela de négocier et de clarifier leurs textes par dialogue avec le système.

Mots clés

Traduction assistée par ordinateur (TAO), Traduction automatique (TA), Traduction humaine assistée par machine (THAM), traduction-dépistage, traduction-diffusion, traduction professionnelle, traduction occasionnelle, stations de travail individuelles pour la traduction.

Introduction

On emploie le terme de « traduction » aussi bien pour la poésie que pour les romans, la publicité, les ouvrages scientifiques, les rapports et manuels techniques, les nomenclatures de pièces détachées, alors qu'il conviendrait, au moins, de distinguer :

- la « re-création », par exemple la traduction d'Edgar Allan Poe par Baudelaire, qui vise avant tout à transmettre l'aspect subjectif, fût-ce au prix d'une légère transformation du contenu;
- la « localisation », largement pratiquée pour les manuels de micro-ordinateurs, qui vise à adapter un contenu à un environnement culturel particulier;
- la « traduction-diffusion » [1.14], en particulier la traduction de documentations techniques dont le contenu doit être strictement rendu, sans ajout ni omission, même si le style « sent la traduction »; et
- la « traduction rapide » enfin, dans laquelle nous rangerons la « traduction-dépistage » de textes écrits et l'interprétation simultanée.

Quelle automatisation de la traduction peut-on souhaiter et réaliser sur des stations de travail individuelles dans les dix prochaines années? Après avoir examiné les types d'automatisation envisageables, et les situations traductionnelles, qui ne se limitent pas à l'image d'Épinal d'une « photocopieuse-traductrice », nous conclurons qu'il vaut mieux en général ne pas parler de « stations de travail », mais plutôt d'environnements ou d'outils, qui ne pourront être spécifiques que dans un cas, celui de la traduction professionnelle en groupe, avec ou sans traduction automatique.

La TAO actuelle

Il n'est absolument pas envisageable pour l'instant d'automatiser la traduction-recréation ni la traduction-localisation plus que par la mise à disposition d'outils d'aide au traducteur humain. Par contre, la « fonction traduisante » est automatisable pour la traduction-diffusion et la traduction-dépistage.

Le terme « TAO » (Traduction assistée par ordinateur) recouvre aujourd'hui l'ensemble des techniques d'automatisation de la traduction. L'ancien terme « TA » (Traduction automatique) est réservé aux techniques d'automatisation de la fonction traduisante, qu'il y ait ou non préédition, postédition ou interaction, tandis que celui de « THAM » (Traduction humaine assistée par la machine ») concerne les outils ou environnements d'aide au traducteur ou au réviseur.

Avant de se demander quelle pourrait être la « station de travail du traducteur » en l'an 2001, il n'est pas inutile de voir où nous en sommes aujourd'hui.

TA pour le dépistage

Vers 1949, les États-Unis, puis l'URSS, ont lancé des programmes de « TA » motivés par le besoin de renseignements. C'est ce que nous appellerons la *TAO pour le veilleur.* Il s'agit de traduction automatique, dont on attend des traductions *grossières*, produites rapidement, en grand volume et à bas coût.

La qualité de ces traductions n'est pas essentielle. Elles servent en effet à filtrer des documents, dont les plus intéressants seront, si nécessaire, traduits ou communiqués à des spécialistes bilingues. Préédition et postédition doivent être absentes ou très limitées (ex. : séparer les phrases, les formules, les figures...).

Les systèmes SYSTRAN sont essentiellement de ce type (par exemple, le système russe-anglais installé depuis 20 ans à la Wright-Patterson Air Force Base traduit, d'après nos informations, environ 18 millions de mots par an, avec une qualité tout à fait satisfaisante pour l'usage visé).

Ce besoin est toujours actuel. Cependant, il s'agit maintenant plus de veille scientifique, technique, économique et financière que de renseignement militaire. À titre d'exemple, on peut citer l'accès en anglais à des bases de données japonaises depuis la Suède [4.6].

TA pour la diffusion

Une quinzaine d'années plus tard, on a commencé à travailler sur la *TAO pour le réviseur.* Il s'agit de produire automatiquement des traductions *brutes*, destinées à être révisées. Dans cette optique, la machine doit remplacer le traducteur, qui est promu réviseur. Cela n'est possible que si l'on restreint convenablement le style et le domaine

des textes à traduire (approche par « sous-optimisation », pour reprendre le terme de L. Bourbeau [1.14]).

Les décideurs (politiques, scientifiques et industriels) comme le grand public n'envisagent souvent que cette possibilité, et ce sans doute à tort. En effet, il existe des systèmes qui peuvent répondre à des besoins de ce type, mais seulement dans des situations convenables. Sinon, l'échec est garanti. Voyons cela un peu plus en détail.

Il existe actuellement près d'une quinzaine de systèmes. Il y a surtout des systèmes japonais (AS-Transac de Toshiba, ATLAS-II de Fujitsu, PIVOT de NEC, HICAT de Hitachi, SHALT d'IBM-Japon, PENSÉE de OKI, MU de l'Université de Kyoto et du JICST...), qui traitent presque uniquement les couples japonais <—> anglais. On peut encore citer des systèmes américains (LOGOS, METAL) ou français (Ariane/aéro/F-E de SITE-B'VITAL, fondé sur les outils informatiques et les méthodes linguistiques du GETA) centrés sur l'anglais, l'allemand ou le français, bien que des maquettes ou des prototypes existent sur de nombreuses autres langues.

Que peut-on espérer de tels systèmes? Essentiellement, de répondre à des besoins de plus en plus importants en traduction technique. Typiquement, en moyenne industrielle, une page de 250 mots est traduite en une heure et révisée en 20 minutes. Dans l'idéal, avec quatre personnes, on passerait donc de trois pages à l'heure à douze pages à l'heure, et on multiplierait donc la productivité par quatre. Il s'agit en fait d'une limite, le chiffre le plus raisonnable étant plutôt de huit pages à l'heure, en comptant une révision plus lourde, de 30 minutes par page, et ce avec des réviseurs formés.

Quand les utiliser? Cela n'est actuellement envisageable que pour de gros flux de textes homogènes et informatisés, comme des manuels d'utilisation ou de maintenance. Dans ces conditions, un système à 1 MF (400 KF de base et 600 KF de spécialisation au vocabulaire et au type de texte) doit pouvoir être amorti en deux ans, pour un flux de 10 000 pages par année (en comptant 10 %/an de maintenance, 60 F/page de coût machine, et 100 F/page de révision, contre 150 F/page de traduction et 70 F/page de révision pour la méthode manuelle classique, soit 60 F/page de gain pour amortir 1,2 MF).

Comment les utiliser? Une condition essentielle de succès de ce type de TAO est de constituer une équipe de développement et de maintenance des linguiciels (dictionnaires, grammaires) qui soit en liaison constante avec l'équipe de révision, et si possible avec les auteurs des documents à traduire. C'est ce qu'a su réussir la PAHO (Pan American Health Organization) [4.7], avec ses systèmes ENGSPAN et SPANAM.

Dans le « contre-rapport ALPAC » du JEIDA [1.12] comme au MTS-II à Munich en août 1989 par exemple, Fujitsu a clairement reconnu avoir fait une erreur en distribuant largement ATLAS-II : seules sont en effet rentables les traductions effectuées chez Fujitsu, soit pour sa documentation, soit dans le cadre d'un contrat avec la CEE, ce dernier ne demandant qu'une révision minimale, car il s'agit en fait de veille technologique. Peut-être est-il approprié de faire ici un parallèle avec les systèmes experts, qui peuvent être développés par des tierces parties, mais qui doivent ensuite être totalement maîtrisés par leurs utilisateurs, seuls à même de les faire évoluer de façon adéquate.

Outils pour traducteurs et réviseurs (THAM)

Que faire pour la plus grande partie des textes dont on veut obtenir de bonnes traductions? La bureautique a commencé à apporter des solutions, sous forme d'outils de *TAO pour le traducteur*. Il s'agit ici de traduction assistée, et non plus automatique (THAM, ou Traduction humaine assistée par la machine). C'est l'utilisateur qui traduit, en s'aidant de dictionnaires bilingues, de bases terminologiques, de thésaurus, de *bitextes* (textes + traductions), etc., accessibles depuis un traitement de texte, le tout formant un *poste de travail pour le traducteur*, réalisé sur micro-ordinateur ou station de travail. Il s'agit d'outils comme Mercury/Termex™ sur PC, WinTool™ [6.10] sur Macintosh, ou de systèmes complets (Weidner-Bravice, TSS de Alps). Le réviseur peut utiliser le même environnement, ou préférer travailler directement sous l'outil final de PAO (Publication assistée par ordinateur).

Pour la majorité des besoins, et en particulier pour la traduction de manuels d'enseignement dans des pays où la langue nationale ne s'est que récemment affirmée comme support de l'enseignement secondaire et universitaire, la THAM est actuellement la seule voie réaliste. Il en est de même de toutes les traductions scientifiques et techniques de faible volume, ou hétérogènes, voire de grands volumes homogènes trop mal rédigés (résumés avec des phrases de 15 lignes, par exemple) ou non disponibles sous forme magnétique cohérente et sans erreurs.

Adapter un système de TAO du réviseur à des besoins comparativement ponctuels serait comme réoutiller une usine pour produire quelques dizaines de voitures. En effet, sans compter la saisie optique ou manuelle, entraînant toujours un coût important de vérification, ni la maintenance, ni même l'achat du système de base, mais seulement sa spécialisation (600 KF) et les coûts de traduction et de révision, on arrive à 632, 680, 760 et 920 KF pour 200, 500, 1000 et 2000 pages, contre 44, 110, 220 et 440 KF pour la méthode classique manuelle, soit environ 14,5, 6, 3,5 et 2 fois plus, respectivement. Il faut aussi tenir compte d'un délai de l'ordre de plusieurs mois pour la spécialisation. Tout compte fait, le point d'équilibre se situe à 9000 ou 10 000 pages.

Les « stations » et « environnements » actuels sont-ils adaptés?

Tentons de répondre à cette question en examinant ce qui existe pour la TA-dépistage, la TA-diffusion et la THAM. Nous verrons qu'il conviendra, pour la suite, de nous intéresser d'abord aux utilisateurs, et ensuite aux techniques.

La TA-dépistage

Accès à un serveur

En France, Gachot SA commercialise un serveur de traduction automatique, via le Minitel, qui donne accès à un certain nombre de « paires de langues » de la famille

des systèmes SYSTRAN. Malheureusement, la publicité cherche à faire croire qu'on peut utiliser ce système en thème, pour la diffusion, alors que les résultats sont évidemment désastreux. Ils le seraient d'ailleurs tout autant avec n'importe quel autre système, car aucun traducteur humain, et *a fortiori* aucun système automatique, ne peut traduire n'importe quoi de façon assez convenable pour une diffusion directe, ni même pour une diffusion après révision (car les réviseurs n'acceptent pas de réviser de la trop mauvaise qualité, et préfèrent retraduire, ce qui, au total, est plutôt contreproductif).

Il reste que, pour le dépistage, et donc en version, l'idée du Minitel est remarquable. Il faudrait seulement la compléter en donnant accès à des bases textuelles en langue étrangère. Cette dernière possibilité a été réalisée en Suède [4.6], où l'on peut accéder à des bases documentaires scientifiques et techniques japonaises, via un PC et un modem, la traduction étant encore réalisée par une version japonais-anglais de SYSTRAN.

Enfonçons le clou : pour ce genre de TA, bien que la qualité de la traduction puisse être jugée extrêmement basse par des traducteurs ou des linguistes, les utilisateurs reconstituent assez aisément le sens de ce qu'ils lisent, et c'est la seule chose qui les intéresse.

À la CEE, SYSTRAN est aussi « utilisé » depuis fin 1976. En fait, il a fallu plus de 12 ans aux services de la Commission pour arriver à la conclusion précédente, c'est-à-dire pour renoncer à faire réviser par des professionnels les sorties de SYSTRAN comme des traductions brutes, et pour les proposer directement aux utilisateurs finals comme des traductions grossières. Selon des informations données par M. L. Rolling fin 1990, on est ainsi passé d'un maximum de 2000 pages (traduites automatiquement et révisées) sur 800 000 pages traduites à la CCE (en 1988), soit 2,5 % du total, à près de 40 000 pages (transmises sans révision) sur 900 000 (en 1989), ce qui reste encore faible dans l'absolu (4,4 %), mais montre bien que, pour un tel flux de textes très variés, seule la TA-dépistage est actuellement viable.

En ce qui concerne les « stations de travail » destinées aux réviseurs, on est passé en 12 ans de terminaux 3270 aux mêmes outils que ceux utilisés par les traducteurs, à savoir des micro-ordinateurs Wang, augmentés d'outils de connexion, ainsi que d'utilitaires liés à la saisie optique. Sans doute les lecteurs actuels des traductions grossières utilisent-ils le même matériel, ou leur micro préféré, ou... le papier.

Stations intégrées

Les matériels deviennent assez puissants pour qu'on puisse mettre sur un micro-ordinateur portable un système de TA couplé avec un système de lecture optique, et un éditeur bilingue. C'est ce que proposent plusieurs sociétés japonaises, comme Toshiba et Sharp. Il s'agit tout de même de machines à 15 ou 17 Mips, avec 8 ou 16 Mo de mémoire centrale, et 180 Mo de disque, sous Unix, et coûtant au moins 15 000 $, sans compter le lecteur optique ni le système.

Le résultat est tout à fait impressionnant. On choisit un fragment d'une page (en anglais), on le numérise, on corrige quelques erreurs de saisie, et ça traduit. Là encore,

la traduction japonaise est très loin d'être parfaite, mais suffit, nous dit-on, à comprendre l'original.

Problème de saisie

Il est clair que l'entrée du texte suppose de la part de l'utilisateur une connaissance minimale du système d'écriture de la langue source, ici l'anglais. C'est le cas de tous les utilisateurs potentiels japonais, dont d'ailleurs une bonne partie connaît suffisamment l'anglais pour ne pas avoir besoin de ce type de système. Que se passerait-il pour saisir des textes en russe ou en coréen (sans parler de l'arabe, qui résiste pour l'instant à la lecture optique)? Ou encore, comment inverser ces systèmes, à l'usage d'Occidentaux désireux de lire en anglais des textes japonais?

Il est vrai que la lecture optique du japonais, fondée sur de gros dictionnaires, donne d'excellents résultats, et que celle des écritures alphabétiques (sauf l'écriture arabe et ses dérivés) devrait arriver à des résultats analogues (moins de 1 % d'erreurs sur les mots, et non sur les caractères), en reprenant le même type de technique. Cependant, les meilleures techniques laissent encore une dizaine d'erreurs par page, ce qui, sans correction, diminue considérablement la qualité que peut produire un système de TA.

TA-diffusion

Éditeur adapté sur grand système

Les systèmes de TA tournant en général sur de gros systèmes, on a commencé par adapter les éditeurs standard, en ouvrant automatiquement deux fenêtres (texte source, texte cible), et parfois trois (dictionnaire classique). Les firmes japonaises ont développé des éditeurs spécifiques, où les phrases originales et traduites (qui restent encore aujourd'hui les unités de traduction, sauf dans les systèmes écrits en Ariane!) sont mises en regard.

Depuis l'arrivée de micro-ordinateurs puissants, munis d'éditeurs et de « texteurs » bien plus conviviaux, cette voie est de moins en moins suivie.

Micro-ordinateur lié à un serveur

En général, le réviseur utilise le même environnement que pour de la traduction-révision usuelle, à savoir un micro-ordinateur classique, connecté au serveur de TA.

Par exemple, le système METAL de Siemens tourne sur une machine LISP, la révision se faisant sur un PC de Siemens. De même, les systèmes ATLAS II de Fujitsu, PIVOT de Nec, HICAT de Hitachi, quand ils sont utilisés en interne pour traduire les documents techniques japonais en anglais, tournent sur de grosses machines, la révision se faisant sur des micros connectés.

De même, SITE-B'VITAL fait tourner le système Ariane/aéro/F-E (sous Ariane-G5) sur gros IBM, à partir d'un micro de base (PC sous MS-DOS avec Word), depuis lequel on a aussi accès à la terminologie spécialisée.

Évolution vers l'intégration?

Les matériels deviennent assez puissants pour tout intégrer, comme nous l'avons vu plus haut. Cependant, cette solution ne semble pas la meilleure pour la TA-diffusion, dans la mesure où l'on vise de grands flux, et où la révision doit donc être réalisée par un groupe de traducteurs.

Par exemple, le système METEO [4.3] tourne depuis longtemps sur micro, mais J. Chandioux a conservé l'architecture avec serveur de TA et révision sur des postes individuels.

THAM

Matériel spécifique et un peu de TA : l'échec est garanti!

C'est la THAM qui a créé le concept de « station du traducteur », et... ses dangers. Bien des déboires sont en effet dus à cette idée qu'il faudrait au « traducteur », sans qu'on sache d'ailleurs s'il travaille en indépendant ou en groupe, un environnement matériel et logiciel spécifique.

Ainsi, ALPS et Weidner ont longtemps proposé du matériel spécifique. Celui d'ALPS était d'origine Convergent Technology, tout comme le Questar-400 que SG2 utilisa, avec aussi peu de bonheur, au temps du PN-TAO [4.2], comme station de révision et poste d'indexage pour un système de TA écrit en Ariane-78. Avec ce type de solution, on arrive toujours à des coûts prohibitifs, à des difficultés de connexion, et à une rapide obsolescence technologique, pour la simple raison que les matériels standard, vendus en très grand nombre, peuvent faire l'objet d'investissements très importants, et deviennent plus puissants. C'est aussi, dans un autre domaine, ce qui est arrivé à Xerox avec ses stations de travail d'IA (1165).

L'idée de proposer un éditeur spécifique est tout aussi mauvaise. Au départ, ceux d'ALPS et de Weidner, pour reprendre les mêmes exemples, étaient fort bons. Mais la grande masse des traducteurs, formée d'indépendants, n'a aucune envie d'apprendre un éditeur ou un texteur nouveau, et à coup sûr très vite dépassé par les grands du marché, comme Word, WordPerfect, etc. C'est aussi l'amère expérience qu'a faite A. Melby, avec son premier système TAIM [6.1], intimement lié au texteur Palantir, qui ne connut pas le succès escompté.

Le pire arrive quand on veut ajouter à ces stations de THAM « un peu de TA », comme l'ont fait ALSP et Weidner. En effet, on se trouve contraint à sur-simplifier le traitement linguistique, pour qu'il tourne en temps raisonnable, et surtout pour que

l'utilisateur puisse lui-même modifier les dictionnaires. En 1985, par exemple, le Bureau des traductions du Secrétariat d'État à Ottawa fit une étude sur Weidner, en demandant aux mêmes traducteurs de traduire le même ouvrage (sur l'odontologie), une fois avec la fonction de TA, et une fois sans, avec un intervalle de temps de deux ou trois mois. Avec la TA, on perdit près de 40 % de productivité... Là comme ailleurs, le mieux est l'ennemi du bien.

Ceci expliquant cela, Weidner a fait faillite, et ALPS s'est reconverti en bureau de traduction, et à la sous-traitance pour IBM, avec TSS (*cf. infra*).

Outils de THAM pure : trois cibles

Il semble y avoir actuellement trois tendances en THAM. D'une part, on trouve des *environnements spécifiques destinés à des professionnels travaillant en groupe*, sur de grosses documentations techniques. C'est là que nous classerons le TSS (Translation Support Service) initialement commandé à ALPS par IBM pour VM/CMS (sur grosses machines) et pour OS/2. L'idée essentielle semble d'être d'intégrer la traduction à l'environnement de documentation technique.

En effet, on peut ainsi considérablement réduire les coûts et augmenter la qualité. D'abord, il est très fréquent qu'on traduise de nouvelles versions d'anciens documents, alors qu'une meilleure gestion permettrait de ne retraduire que les passages qui ont effectivement changé. Ensuite, l'accès aux traductions passées peut fournir rapidement la solution à de nombreux problèmes de traduction, et augmenter l'homogénéité de l'ensemble des traductions.

D'autre part, on trouve des *outils destinés aux professionnels indépendants*. Il s'agit surtout de programmes résidant sur PC (comme Mercury/Termex [6.5]) ou d'accessoires de bureau sur Macintosh (comme WinTool), qui permettent d'accéder à un ou plusieurs dictionnaires « en ligne », depuis n'importe quelle application. Il est très important que l'utilisateur puisse définir des dictionnaires personnels et les modifier.

On trouve enfin des *environnements destinés à des traducteurs occcasionnels*, comme SISKEP [6.6, 6.8], déjà bien diffusé en Malaisie. Il s'agit pour l'instant de versions préliminaires, diffusées gratuitement, mais des produits devraient suivre. Là, le traducteur n'est pas un professionnel, et ne traduit pas nécessairement vers sa langue maternelle. Même si c'est le cas, il ne connaît souvent pas certains termes spécifiques, les ayant appris dans la langue source (cas de l'anglais-malais, du français-arabe, par exemple).

La traduction occasionnelle requiert ainsi, outre des dictionnaires bilingues en ligne, des outils liés à la rédaction dans la langue cible, comme des correcteurs d'orthographe, des « critiqueurs » de style, des conjugueurs, des thesaurus, etc.

Peut-être à cause de la simplicité morphologique de l'anglais, SISKEP est le seul système qui effectue une lemmatisation de la chaîne sélectionnée en cas d'échec de la recherche directe dans le dictionnaire, pour éviter à l'utilisateur d'entrer lui-même la forme canonique. Il s'agit d'une propriété très utile, que nous avons cherché à généraliser en concevant notre maquette ODILE [6.9] sur Macintosh.

Propositions, maquettes, idées...

Il y a eu ces dernières années plusieurs études sur la « station du traducteur », avec des maquettes, comme celles du CCRIT au Canada et de CAP-Gemini-Innovation en France. Le sujet est certainement actuel. Ces études renforcent l'idée que les outils de THAM doivent être les moins spécifiques possible, et surtout qu'il faut proscrire tout matériel spécifique.

L'un des problèmes cruciaux pour les traducteurs reste l'absence d'environnements vraiment multilingues. Bien sûr, il existe des adaptations de logiciels à telle ou telle langue, et même, sur le Macintosh, de nombreuses variantes du système d'exploitation, adaptées aux langues de grande diffusion. Il ne s'agit toutefois que de « localisation » : on traduit les messages et les formats particuliers (nombres, dates, prix), et on traite un système d'écriture particulier (arabe, chinois, japonais...), en plus du système de base, fondé sur l'alphabet romain. Mais il reste impossible de fabriquer une nomenclature multilingue comportant, par exemple, du japonais et de l'arabe.

Quatre grandes situations pour le futur

Le concept de « station du traducteur » nous semble trop limité à la traduction professionnelle. Or, les utilisateurs de TAO seront de plus en plus des non-professionnels ou des amateurs. Selon la compétence de l'utilisateur dans les langues visées, et, pour le professionnel, selon qu'il travaille en indépendant ou en groupe, nous pouvons distinguer quatre types de situations pour le futur.

Traduction « dépistage » individuelle

Les serveurs de TA et les solutions intégrées devraient coexister.

Accès à un serveur

Les serveurs paraissent appropriés pour toutes les situations où une information peut être demandée par un grand nombre de personnes. Ainsi, on ne la traduira qu'une fois, et éventuellement à l'avance. Cela permet également d'avoir des versions spécialisées par type de textes, et d'effectuer dans certains cas un « toilettage » très rapide des sorties (comme le fait actuellement Fujitsu sur des traductions grossières fournies à la CCE).

On retrouve ici les situations classiques, où l'information est déjà numérisée : accès à des bases de données documentaires, à des messages éphémères (comme la météo ou les dépêches d'agence). On peut aussi penser à la « lecture active » dans les bibliothèques informatisées du futur.

On peut imaginer un service d'aide à la saisie, dans le cas où le lecteur n'a aucune connaissance du système d'écriture de la langue source et où la lecture optique laisse

à désirer, par manque de qualité de l'original ou toute autre raison. Si l'utilisateur a un lecteur optique, l'aide aurait lieu à distance. Sinon, il se déplacerait ou enverrait l'original.

Solution intégrée sur gros micro

Il s'agit, encore plus que dans le cas précédent, de systèmes à large spectre. On imagine des utilisateurs documentalistes dans de grandes organisations ou firmes, cherchant à savoir ce qui se fait dans tel ou tel domaine, et ce dans plusieurs langues.

Soit l'information est déjà numérisée (disquettes, bandes, DOC, bases de données non munies de serveurs de TA...), et il n'y a « que » des problèmes de traduction de formats. Soit elle ne l'est pas, et l'utilisateur doit acquérir une connaissance minimale du ou des systèmes d'écriture utilisés, ce qui reste tout à fait possible dans le cadre évoqué. Par exemple, il ne faut pas une longue formation pour répondre à un système qui vous demande quel caractère chinois choisir, parmi une liste d'une dizaine.

Vu la tendance actuelle, on risque peu de se tromper en disant qu'il s'agira de vraies « stations de travail », au sens où on l'entend aujourd'hui (10 à 20 Mips, Unix, X-Window).

Nécessité d'environnements génériques (matériel, logiciel)

Dans un cas comme dans l'autre, les environnements seront génériques. La seule différence concernera la puissance requise. Pour l'accès à un serveur, un micro de base suffit et suffira. Pour les solutions intégrées, il faudra une station de travail standard.

Dans le premier cas, il est évident qu'on ne peut imposer un logiciel spécifique pour l'édition des résultats. Il faudra fournir des « filtres » (transformateurs de formats) pour les standards du marché. Dans le second cas, on ne le peut pas non plus, car il est fort probable que l'utilisateur devra mettre plusieurs systèmes sur sa station, ayant à traduire à partir de plusieurs langues sources. En effet, pour les constructeurs de systèmes de TA, il est plus rentable et plus facile de travailler à partir d'une ou deux langues vers de nombreuses langues que l'inverse.

Traduction « occasionnelle »

Il s'agit de personnes qui connaissent bien les deux langues, mais pas comme des traducteurs, et qui ne traduisent pas nécessairement vers leur langue maternelle, ni même à partir d'elle. En Malaisie, par exemple, il est fréquent que des Chinois rédigent des notes de cours en anglais et les traduisent ensuite en malais. En Afrique du Nord, de nombreux scientifiques formés en France peuvent avoir à traduire vers l'arabe à partir de l'anglais aussi bien que du français.

Dans ce contexte, il faut absolument des outils simples, tournant sur des micros de bas de gamme, si possible en conjonction avec la plupart des applications du commerce (tableurs, texteurs, grapheurs, SGBD, documenteurs...).

Étant donné les progrès techniques et la baisse rapide des prix, on peut penser trouver, sur de tels matériels, les fonctions actuelles (recherche dans les dictionnaires bilingues avec lemmatisation automatique, thesaurus en langue cible, fonctions de manipulation morphologique en langue cible), présentées de façon bien plus ergonomique, c'est-à-dire au moyen de fenêtres « vocables » qui seraient liées à la fenêtre de l'application (texteur par exemple), et contiendraient en permanence les informations associées aux termes de cette fenêtre. Le travail de recherche se faisant en tâche de fond, l'utilisateur n'aurait plus à sélectionner, demander une recherche, attendre, etc.

Traduction professionnelle « individuelle »

Le traducteur professionnel traduit en principe dans sa langue, ou dans une langue qu'il connaît parfaitement, ainsi que la plupart des termes, généraux et techniques, de ce qu'il doit traduire. Par conséquent, si les outils précédents peuvent lui être utiles, des outils de recherche intelligente dans ses traductions passées lui seraient bien plus profitables.

Des outils de communication sont également indispensables, pour la consultation de bases terminologiques dans les cas particulièrement délicats. Même si on a un DOC, aucune base terminologique n'est jamais complète.

Par conséquent, on voit bien un traducteur indépendant s'équiper d'un micro « classique », plutôt de haut de gamme (AT-386 sous Windows, ou Mac LC, par exemple), avec beaucoup de mémoire sur disque, un lecteur de DOC et des outils de connexion.

Verra-t-on des éditeurs ou texteurs spécialement conçus pour les traducteurs indépendants? C'est fort douteux, vu que ce marché est étroit, que les clients demandent les formats usuels, et que les texteurs offriront sans doute la plus grande partie des fonctionnalités souhaitables, comme la possibilité de synchroniser deux ou plusieurs fenêtres (comme WinText le fait déjà).

Traduction professionnelle « industrielle »

La traduction professionnelle de haute qualité est souvent produite par plusieurs personnes. Habituellement, le traducteur qui effectue le premier jet a une compétence technique très superficielle, mais connaît très bien la terminologie et les deux langues. Le premier réviseur est un traducteur « senior », spécialiste du type de document considéré, et à même d'assurer l'homogénéité terminologique et stylistique. Enfin, on fait parfois intervenir un second réviseur, spécialiste du contenu technique du document et éventuellement ignorant de la langue source, pour détecter les contresens sémantiques

linguistiquement plausibles, ainsi que d'éventuelles ambiguïtés accidentellement introduites en langue cible.

D'autre part, les documents à traduire sont souvent très longs, et doivent être traduits très rapidement. Il faut donc synchroniser le travail de plusieurs traducteurs et réviseurs. Ainsi, la traduction professionnelle industrielle justifie-t-elle la conception d'environnements plus spécifiques que la traduction professionnelle individuelle.

L'introduction de la TA dans ce type d'activité se généralisera sans doute. Aux fonctionnalités mentionnées ci-dessus, il faudra donc rajouter celles qui ont trait à l'amélioration du système de TA. Pour cela, il faut pouvoir recenser les corrections des réviseurs, leur permettre d'envoyer des messages aux développeurs-mainteneurs du système, et éventuellement leur donner accès, sous une forme appropriée, aux lexiques du système, pour leur permettre de tenter un diagnostic, et en tout état de cause supprimer l'aspect « boîte noire », psychologiquement très néfaste.

La fonction de TA étant réalisée par un serveur, pour les raisons indiquées plus haut, il n'est pas nécessaire de viser des stations de travail. Le type de matériel nécessaire sera plutôt fonction de la puissance requise par l'ensemble des outils liés à la gestion de la documentation technique et de la traduction, s'il s'agit de micros. Mais il pourra tout aussi bien s'agir de « terminaux intelligents », successeurs des « terminaux X » actuels, connectés à un serveur de gestion de la documentation, de la traduction et des tâches de TA.

Perspectives et conclusions

Nous avons évoqué quatre types d'utilisateurs, le veilleur, le traducteur occasionnel, le traducteur professionnel indépendant et le traducteur professionnel industriel. Seuls le premier et le quatrième peuvent à notre avis utiliser avec profit les techniques de TA existantes. Il reste encore une possibilité, actuellement au stade des recherches préliminaires, celle de la TAO personnelle [5.8] ou *TAO pour le rédacteur.*

De plus en plus, nous désirons rédiger dans notre langue, et transmettre nos textes à l'étranger, qu'il s'agisse de messages électroniques, de lettres, d'articles (on pense à la récente controverse sur l'Institut Pasteur), de manuels techniques, voire de livres. En effet, contrairement à ce que d'aucuns prédisaient il y a une cinquantaine d'années, l'internationalisation croissante ne s'est pas accompagnée d'une uniformisation linguistique vers l'anglais, mais au contraire d'un renforcement considérable de l'usage scientifique et technique des langues traditionnellement importantes de ce point de vue, et d'une promotion volontariste de bien d'autres, pour les amener au même niveau (malais-indonésien ou arabe, par exemple). À notre sens, cette évolution ne fera que se renforcer, les langues étant, comme le note Claude Hagège dans un article paru dans *Le Monde* début 1990, les *drapeaux des identités nationales.*

Au GETA, nous sommes en train d'étudier et de maquetter ce nouveau concept de « TAO personnelle » dans le cadre du projet LIDIA. Une hypothèse de base est que le dialogue nécessaire, même s'il est lourd (de l'ordre de grandeur de la traduction-révision humaine vers une seule langue cible), serait acceptable dans la mesure où

l'auteur, et non le système, le déclencherait, où une réalisation asynchrone éliminerait toute attente forcée, et où le système serait « contrôlable » et « découvrable » par l'utilisateur, grâce à des possibilités de réglage de paramètres, d'intervention au niveau des « marques cachées », de « rétrotraduction », et d'exploration des connaissances lexicales et grammaticales contenues dans le système, présentées de façon « naturelle » et non codée, ce qui est d'ailleurs un problème en soi.

Les trois types de TAO *classique* ne conviennent évidemment pas à ce nouveau besoin. En effet, la TAO du veilleur (TA-dépistage), sans préédition ni postédition, ne peut donner une qualité suffisante, et la TAO du réviseur (TA-diffusion) comme la TAO du traducteur (THAM) s'adressent par définition à des spécialistes au moins bilingues, et non à des rédacteurs du grand public, supposés ne connaître aucune des langues cibles (ou au plus une, et ce imparfaitement).

La « station du rédacteur », dans un système de TAO personnelle, devrait, pour des raisons de diffusion, être un micro-ordinateur répandu, de bas ou de milieu de gamme. Elle supporterait les logiciels de vérification orthographique, terminologique et stylistique, de dialogue, et de communication, tandis que le système de TA tournerait sur un serveur (connecté à un réseau local ou accessible par Minitel). Du point de vue logiciel, il faudra utiliser un outil programmable du marché. Aucun texteur ne l'étant pour l'instant, nous avons choisi de maquetter avec HyperCard, qui présente d'ailleurs d'autres avantages, tant ergonomiques que linguistiques [5.5]. Mais cela évoluera sans doute. En tout état de cause, il y a encore beaucoup de recherche à faire sur ce concept, puis d'importants investissements linguistiques, avant de voir des produits de TAO personnelle. Espérons que ce sera une perspective proche en l'an 2001 !

Références (thématiques et chronologiques)

1. Références générales sur la TAO

[1.1] KITTREDGE R. (1983) « Sublanguage — Specific Computer Aids to Translation — a Survey of the Most Promising Application Areas », *Contract* n° 2-5273, Montréal, Université de Montréal et Bureau des traductions, mars 1983, 95 p.

[1.2] BOITET Ch. (1985) *Traduction (assistée) par ordinateur : ingénierie logicielle et linguicielle*, Conf. AFCET RF&IA, Grenoble, nov. 1985.

[1.3] HUTCHINS W. J. (1986) *Machine Translation : Past, Present, Future*, Chichester, England, Ellis Horwood, John Wiley & Sons, 382 p.

[1.4] KITTREDGE R. (1986) *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*, Grishman R. & Kittredge R. (eds.), Hillsdale, New-Jersey, Lawrence Erlbaum.

[1.5] DESCLÉS J. P. (1987) *Technologos*, Paris, LISH-CNRS, printemps 1987.

[1.6] BOITET Ch. (1988) « Software and Lingware Engineering in Modern M(A)T Systems », Bátori (ed.) *Handbook for Machine Translation*, Tübingen, Niemeyer.

[1.7] BOITET Ch. (1988) « PROS and CONS of the Pivot and Transfer Approaches in Multilingual Machine Translation », *New Directions in MT*, BSO conf, Budapest, 13 p.

[1.8] BOITET Ch. (1988) *L'apport de Bernard Vauquois à la traduction automatique et au traitement automatique des langues naturelles*, Colloque sur l'Histoire de l'informatique en France, Grenoble, 3-5 mai 1988, vol. 2, p. 63-82.

[1.9] LEHRBERGER J. et BOURBEAU L. (1988) « Machine Translation. Linguistic Characteristics of MT Systems and General Methodology of Evaluation », The Hague, John Benjamins, *Lingvisticæ Investigationes*, 15, 240 p.

[1.10] VAUQUOIS B. (1988) *BERNARD VAUQUOIS et la TAO, vingt-cinq ans de traduction automatique, ANALECTES. BERNARD VAUQUOIS and MT, twenty-five years of MT*, Ch. Boitet (éd.), Grenoble, Ass. Champollion & GETA, 700 p.

[1.11] ABBOU A., (éd.) (1989) *Traduction assistée par ordinateur*, Actes du séminaire international (Paris, 17-18 mars 1988) et dossiers complémentaires, Paris, DAICADIF, Observatoire des Industries de la langue, 234 p.

[1.12] JEIDA (1989) *A Japanese View of Machine Translation in Light of the Considerations and Recommendations Reported by ALPAC, USA*, Tokyo, Japanese Electronic Industry Development Association, 197 p.

[1.13] BOITET Ch. (1990) « Multilingual Machine Translation Does Not Have to Be Saved by Interlingua », *Proc. MMT'90*, Tokyo, 5-6 novembre 1990.

[1.14] BOURBEAU L. (1990) *Élaboration et mise au point d'une méthodologie d'évaluation linguistique de systèmes de Traduction Assistée par Ordinateur*, rapport final, Ottawa, Secrétariat d'État du Canada, mars 1990, 203 p.

[1.15] ROLLING L. (1990) « Trends of Multilingual Machine Translation in Europe », *Proc. MMT'90*, Tokyo, 5-6 novembre 1990, 2 p.

2. Compréhension et traduction (humaine, automatique)

[2.1] BAR-HILLEL Y. (1960) « The Present Status of Automatic Translation of Languages », F.L. Alt (ed.), *Advances in Computers (volume 1)*, New York, Academic Press, p. 91-163.

[2.2] BAR-HILLEL Y. (1964) *Language & Information*, Addison Wesley.

[2.3] BAR-HILLEL Y. (1971) « Some Reflections on the Present Outlook for High Quality Machine Translation », *Feasibility Study on Fully Automatic High Quality Machine Translation*, RADC-TR-71-295, Austin, Univ. of Texas.

[2.4] NIRENBURG S. et GOODMAN K. (1990) « Treatment of Meaning in MT Systems », *Proc. ROCLing-III*, Taipeh, 20-22 août 1990, p. 83-101.

3. TA « fondée sur le contenu »

[3.1] WILKS Y. (1973) « An Artificial Intelligence approach to Machine Translation », Shank & Colby (eds.), *Computer Models of Thought and Language*, Freeman & Co, p. 114-151.

[3.2] BOITET Ch. et GERBER R. (1986) « Expert Systems and Other New Techniques in MT », Bátori & Weber (eds.), *Neue Ansätze in maschineller Sprachübersetzung*, Tübingen, Niemeyer, p. 103-119.

[3.3] BOITET Ch. (1988) « Representation and Computation of Units of Translation for Machine Interpretation of Spoken Texts », *TR-I-0035*, ATR, Osaka, 41 p. Also *Computers & Artificial Intelligence 8 /6*, p. 505-546, 1989.

[3.4] NIRENBURG *et al.* (1989) *KBMT-89 Project Report*, Pittsburgh, Center for Machine Translation, Carnegie Mellon University, February 1989, 286 p.

[3.5] NIRENBURG S. (1989) « Knowledge-Based Machine Translation », *Machine Translation **4/1***, March 1989, 5-24.

[3.6] NIRENBURG S. et DEFRISE Ch. (1990) « Lexical and Conceptual Structure for Knowledge-Based Machine Translation », *Proc. ROCLing-III*, Taipeh, 20-22 août 1990, p. 105-130.

[3.7] CARLSON L. et NIRENBURG S. (1991) « World Modeling for NLP », *CMT, CMU*, 31 p.

[3.8] BOITET Ch. (1991) « Un système de traduction automatique peut-il et doit-il comprendre? » Paris, Actes de la Convention IA-91, 15-17 janvier 1991, *Hermès*, p. 13-25.

4. TA « fondée sur la forme »

[4.1] BOITET Ch. et NEDOBÉJKINE N. (1981) « Recent Developments in Russian-French Machine Translation at Grenoble », *Linguistics **19** (1981)*, p. 199-271.

[4.2] BOITET Ch. (1986) « The French National MT Project : Technical Organization and Translation Results of CALLIOPE-AERO », *Computers and Translation **1*** (1986), p. 281-309.

[4.3] CHANDIOUX J. (1988) « 10 ans de METEO CMD) », *In [1.11, Abbou 89]*, p. 169-173.

[4.4] DUCROT J. M. (1988) « Le système TITUS IV », *In [1.11, Abbou 89]*, p. 55-71.

[4.5] PECCOUD F. (1988) « The Aims of the French National Project of Computer-Aided Translation », *International Forum on Information and Documentation*, **13**/1, p. 11-13.

[4.6] SIGURDSON J. et GREATEX R. (1987) « MT of On-Line Searches in Japanese Data Bases », *RPI*, Lund Univ., 124 p.

[4.7] VASCONCELLOS M. et LEÓN M. (1988) « SPANAM and ENGSPAM : Machine Translation at the Pan American Health Organization », J. Slocum (ed.), *Machine Translation Systems*, Cambridge Univ. Press, p. 187-236.

5. TAO « fondée sur le dialogue »

[5.1] RICHARDSON S. D. (1985) *Enhanced Text Critiquing using a Natural Language Parser : the CRITIQUE System*, Yorktown Heights IBM Research Report RC 11332.

[5.2] CHANDLER B., HOLDEN N., HORSFALL H., POLLARD E. et MCGEE WOOD M. (1987) *N-Tran Final Report*, Manchester, Alvey Project, CCL/UMIST Report 87/9, 30 p.

[5.3] ZAJAC R. (1988) « Interactive Translation : a New Approach », *Proc. COLING-88*, Budapest, 22-27 août 1988, p. 785-790.

[5.4] BOITET Ch. (1989) « Motivations and Architecture of the LIDIA Project », *Proc. MTS-89*, Munich, août 1989, 5 p.

[5.5] BOITET Ch. (1989) « TAO Personnelle et HyperTexte », *Actes du colloque BUROTICA-89*, Paris, 10-13 octobre 1989, 6 p.

[5.6] BOITET Ch. (1989) « Speech Synthesis and Dialogue-Based Machine Translation », *Proc. ASTI Conf.*, ATR, Nara, 11-12 décembre 1989, 9 p.

[5.7] BLANCHON H. (1990) *LIDIA-1 : un prototype de TAO personnelle pour rédacteur unilingue*, X^{e} journées internationales sur « les systèmes experts et leurs applications », Avignon, juin 1990, 10 p.

[5.8] BOITET Ch. (1990) « Towards Personal MT : General Design, Dialogue Structure, Potential Role of Speech », *Proc. COLING-90*, Helsinki, 20-25 août 1990, 6 p.

[5.9] BLANC E. & BOITET Ch., (eds) (1990) « DBMT-90, a Workshop on Dialogue-Based Machine Translation : Documents & Slides », *Le Sappey*, 26-28 août 1990, Grenoble, GETA, IMAG, 350 p.

[5.10] BOITET Ch. (1990) « TAO personnelle et promotion des langues nationales : le projet LIDIA du GETA », *Tribune des Industries de la Langue*, Grenoble, automne 1990.

6. Outils de THAM

[6.1] MELBY A. K. (1978) « Design and Implementation of a Computer-Assisted Translation System », *Proc. COLING-78*, Bergen, 14-18 août 1978, 28 p.

[6.2] HUNDT M. G.. (1982) « Working With the Weidner Machine-Aided Translation System », V. Lawson (ed.) *Practical Experience of Machine Translation*, The Hague, ASLIB, North Holland, Conf., p. 45-51.

[6.3] MELBY A. K. (1982) « Multilevel Translation Aids in a Distributed System », *Proc. COLING-82*, Prague, 5-10 juillet, North Holland, p. 215-220.

[6.4] LONSDALE D. (1984) *Notes on Interactive Translation*, ALPS Inc., Provo, Utah, 10 p.

[6.5] MELBY A. K. (1984) *MERCURY ™ Card File and Glossary Manager. User's Guide*, Provo, Utah, LinguaTech Inc., 114 p.

[6.6] TONG L. C. (1986) *SISKEP, a Software for the Human Translator*, PTMK, Penang, Universiti Sains Malaysia, 6 p.

[6.7] KJAERSGAARD P. S. (1987) « REFTEX, un progiciel pour la traduction assistée par ordinateur », *Computers and Translation*, 7 p.

[6.8] TONG L. C. (1987) « The Engineering of a Translator Workstation », *Computers and Translation 2/4*, 263-273.

[6.9] TOMASINO I. (1990) *ODILE, un outil d'intégration extensible de lemmatiseurs et de dictionnaires*, Mémoire CNAM, GETA, Grenoble, décembre 1990, 151 p.

[6.10] WINSOFT (1987) *Manuel d'utilisation de WinTool ™. Version 1.1*, Grenoble, WinSoft.

2

L'environnement sémantique, aide au transfert lexical en TA. Problèmes posés par le calcul des traductions françaises de quelques verbes anglais

Arlette ATTALI, Annette EUVRARD

CELTA, CNRS/EUROTRA F-Nancy, Université de Nancy II, Nancy, France

Nous nous proposons de présenter une synthèse d'un travail réalisé en collaboration avec des membres de l'équipe du Centre d'études linguistiques pour la traduction automatique.

Cette étude s'intègre dans le cadre général des problèmes rencontrés dans tout programme de TA, lequel comporte, nous le rappelons, trois phases : l'**analyse**, le **transfert** et la **génération**. Nous nous sommes plus particulièrement intéressés aux difficultés propres à la **phase transfert**; celle-ci étant le moment où le signifiant de la langue cible se substitue au signifiant de la langue source.

Dans le cas qui nous occupe, c'est-à-dire la traduction bilingue anglais → français, nous avons cherché des moyens linguistiques susceptibles de nous aider à résoudre les problèmes spécifiques liés au **transfert lexical des verbes anglais** en TA[1].

1. *Cf.* publication : « Transfert lexical et contexte sémantique en Traduction Automatique. Les traductions françaises des verbes anglais DRAW, HOLD, LIE, TURN, WORK ». A. Attali, G. Bourquin, M. C. Bourquin-Launay, A. Euvard. CELTA, URA-CNRS 1035. Cahiers du CRAL, Nancy, 1989.

Pour mieux comprendre ces problèmes, nous prenons l'exemple du verbe HOLD, appréhendé dans des textes réels et limités à la phrase, avec une multiplicité de traductions possibles (*cf.* annexe 1). Se pose donc la question de savoir comment un automate-traducteur qui traduit de lexème à lexème peut sélectionner l'équivalent lexical adéquat : chaque traduction française correspondant à un des sens de HOLD doit être l'**aboutissement d'un calcul**. Ce calcul consiste, pour nous, à tirer parti de l'**environnement syntactico-sémantique** porté par les **actants** du verbe anglais. Par actants nous entendons : les sujets profonds ou GN1, les objets directs ou indirects profonds ou GN2 et parfois GN3 auxquels on pourrait ajouter des circonstants plus ou moins attendus.

Le principe est donc de dégager de cet environnement actanciel des informations **minimales** mais aussi **représentatives d'un nombre suffisamment élevé d'exemples** pour pouvoir servir de critères de sélection des traductions dans un programme automatique.

Étant donné la rigidité du système, on est confronté à une double difficulté avec laquelle il faut composer : 1) difficulté du choix de l'équivalent lexical français qui doit couvrir le plus grand nombre de cas possibles et en même temps préserver le sens originel, et 2) difficulté du choix des critères de calcul de cet équivalent lexical, les deux choix étant liés l'un à l'autre.

Nous avons cherché un début de réponse à ces problèmes en commençant par examiner **cinq verbes anglais** : DRAW, HOLD, LIE, TURN, WORK, dans toutes leurs formes grammaticales, à tous leurs temps et accompagnés de toutes leurs particules. Ils ont été relevés dans un corpus de concordances de 3 500 000 occurrences parmi lesquelles ces verbes représentent 5465 attestations.

C'est cette quantité d'exemples tirés de textes authentiques — et non fabriqués — sur laquelle s'appuie notre étude, qui nous permet de répertorier certaines régularités contextuelles afin d'élaborer les règles de calcul applicables au transfert lexical des verbes.

– Nous avons choisi ces cinq verbes pour les raisons suivantes :

1) Ils sont sémantiquement différents les uns des autres.
2) Chacun d'eux a des effets de sens très différents en anglais : concret/abstrait, propre/figuré.
3) Chaque verbe donne lieu à un nombre assez élevé de traductions françaises différentes.
4) Chacun a une fréquence d'emploi suffisamment élevée dans les corpus pour que l'étude soit significative.
5) Enfin, chacun s'intègre dans des réseaux actanciels suffisamment diversifiés et riches.

– La mise en œuvre de l'expérience se décompose en une suite de tâches strictement hiérarchisées. Chaque verbe est examiné cas par cas dans les phrases où il apparaît. Il faut donc :

1) Choisir **empiriquement** la traduction française en fonction du contexte; cette traduction se veut aussi exacte et généralisable que possible.

2) Déterminer le minimum d'information syntactico-sémantique nécessaire au calcul de cette traduction.
3) Élaborer des règles informatisables qui représentent et codifient l'information jugée pertinente.
4) Construire l'algorithme qui rend compte du déroulement hiérarchisé des règles de calcul et qui est représenté sous forme d'organigramme.

Ces différentes étapes se conditionnent l'une l'autre et c'est par approximations successives qu'on parvient finalement au choix de traduction.

Rappelons que l'élaboration de ces règles de traduction suppose qu'un certain nombre de problèmes aient été déjà résolus au cours de la phase d'analyse qui précède, sur le plan syntaxique et sémantique. Par exemple, on considère comme acquises les levées d'ambiguïtés grammaticales (substantif/verbe; adjectif/adverbe); la reconnaissance des référents des pronoms relatifs et personnels, celle des circonstants non prépositionnels et des différentes fonctions syntaxiques en général.

Critères de sélection des traductions fondés sur la morpho-syntaxe

Nous ne donnerons que quelques exemples indicatifs :

1- **Un mot spécifique**; par exemple la particule qui accompagne le verbe :
   – Field **DREW AWAY** on his motorcycle.
   Règle : DRAW + AWAY ⟶ « S'ÉLOIGNER ».

2- **Un élément grammatical** :
   – The observation **HOLDS true** whether...
   Règle : HOLD + Adj. ⟶ « RESTER ».

3- **La voix** à laquelle le verbe est lui-même employé peut être un élément qui oriente vers la bonne traduction :
   – I **HELD** receptions for leading figures
   Règle : HOLD = ACTIF ⟶ « ORGANISER ». (Ici ce n'est pas le seul élément à être pris en compte mais il est nécessaire.)

4- **Une fonction syntaxique** :
   – The culture was **HELD at 34 degrees F**.
   Règle : HOLD + circ. de mesure ⟶ « MAINTENIR ».

Critères de sélection des traductions fondés sur la sémantique :

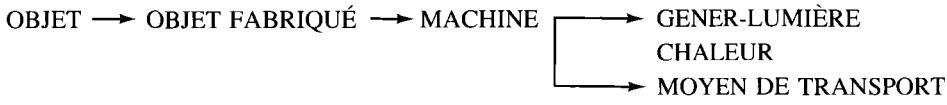
C'est ici qu'interviennent les **descripteurs sémantico-cognitifs** que l'on peut définir comme classes notionnelles regroupant un ensemble d'unités lexicales ayant un trait sémantique commun. Ils se caractérisent de plusieurs manières :

Ces descripteurs ne sont ni de même niveau ni de même nature.

En effet, comme ils ont été établis de manière empirique, **en fonction de la traduction souhaitée**, il y a une grande disparité dans l'étendue de leur champ d'application respectif.

À côté des grands traits sémantiques généraux bien connus tels que HUMAIN, ANIMAL, ABSTRAIT, OBJET, etc., d'autres font référence à des notions très spécifiques : AFFECTIVITÉ, FINANCE, CONNOTATION PÉJORATIVE, PORTEUR D'INFORMATION, PORTEUR DE VIRTUALITÉS, etc. (*cf.* annexe 2).

Ces descripteurs ne sont pas hiérarchisables, tout au plus peut-on en regrouper certains en allant du plus général au plus particulier :



Il convient de souligner que le descripteur ne rend pas compte de la totalité du sémantisme de l'unité lexicale à laquelle il est affecté; il ne retient qu'un des aspects couvrant un champ plus ou moins large. C'est ainsi que certains lexèmes peuvent être classés sous plusieurs descripteurs différents. Un cas intéressant est par exemple le mot « **POSITION** » classé sous quatre descripteurs différents et « **BOOK** » sous trois (*cf.* annexe 3).

Ces descripteurs sont de rentabilité variable.

Un descripteur est « rentable » lorsqu'il est fréquemment utilisé : soit parce qu'il recouvre un grand nombre de termes, soit parce qu'il est utilisé pour plusieurs verbes, soit parce qu'il réalise les deux choses à la fois.

C'est le cas pour le descripteur **HUMAIN**, statistiquement le plus rentable : il est utilisé pour les cinq verbes, il recouvre un très grand nombre de mots et il est appliqué aux trois fonctions GN1, GN2, GN3.

Par contre le descripteur « **PHÉNOMÈNE SOCIAL TEMPORAIRE** » ne concerne que le verbe HOLD, mais s'applique à un très grand nombre de termes. De plus il est sémantiquement bien délimitable (*cf.* annexe 4).

Le descripteur « **ENTITÉ SPATIALE, GÉOGRAPHIQUE, GÉOLOGIQUE** » concerne trois verbes: DRAW, HOLD et LIE; mais comme il est très général, la limite de son extension est difficile à établir (*cf.* annexe 4).

Enfin, on a créé des descripteurs, pour le moment très faiblement rentables, comme « **TEMPS** », « **BRUIT** », « **PHÉNOMÈNE NATUREL** » (tels que « WAVE LIFT ») qui ne concernent chacun qu'un seul verbe, ne recouvrent qu'un seul mot lui-même attesté dans un seul exemple. Cela est une exception, car nous n'avons pas voulu faire des traductions *ad hoc*; mais dans ces cas on a pensé que

ces descripteurs étaient susceptibles d'accroître leur rendement au fur et à mesure qu'augmenterait le nombre des verbes étudiés. L'expérience montrera si cette intuition était fondée ou non.

Application des descripteurs aux actants concernés :

Les exemples suivants (tirés de l'annexe 1) mettent en évidence le rôle de la sémantique actancielle pour aider au transfert lexical des verbes dans un programme de TA.

1. The **senate seat** *held* by her husband Richard.
   GN2
   [socio-prof. abstrait] **occuper**
4. The century **dinner** *was held* at Eton.
   GN2
   [phén. social temp.] **avoir lieu**
11. The **incubator** *holds* yoghourt jars.
    GN1
    [contenant] **contenir**
14. Party **leaders** *holding* **shares** in private firms...
    GN1 **GN2**
    [humain] **détenir** **[finance]**
15. A surplus of labor *holds back* the **mechanization** of...
    GN2
    freiner **[processus]**

Limites liées au choix des descripteurs sémantiques :

Dans l'idéal, les règles édictées par l'algorithme de traduction devraient permettre de traduire avec exactitude la totalité des attestations du corpus utilisé. Mais ce n'est pas le cas.

Aussi à côté des traductions jugées « **satisfaisantes** » (qui représentent tout de même environ 85 % des cas dans notre corpus), nous avons des traductions dites « **approximatives** » (7 % — il s'agit des cas où la plus grande partie du sens source est conservée), des traductions « **impropres** » (3 % — cas où la plus grande partie du sens source est perdu, mais reste compréhensible grâce au contexte), et enfin des traductions « **incorrectes** » (3 % — cas où l'on a perdu totalement et le sens source et la compréhension).

Nous allons illustrer les TROIS derniers cas (c'est-à-dire les traductions non satisfaisantes) en essayant d'analyser les causes d'imprécisions ou d'erreurs.

– **Première cause** de traductions non satisfaisantes : le fait que nous n'avons pas voulu produire des descripteurs entièrement *ad hoc* afin de ne pas multiplier les traductions uniques (nous ne l'avons fait que pour trois ou quatre cas bien particuliers dont nous avons parlé plus haut). Cela entraîne une traduction par défaut, inadaptée.

Exemple :

The relation theory does not **HOLD UP** very well.

Règle suivie : HOLD + UP intr. ⟶ « SE MAINTENIR » au lieu de « RÉSISTER ».

Il aurait fallu créer un trait sémantique sous le GN1 « theory » pour ce seul cas.

Autre exemple :

A wide-ranging force [of Republicans]... was **DRAWN UP** against a solid phalanx of southern Democrats.

Règle suivie : DRAW + UP + GN2 ⟶ traduction par défaut ⟶ « ÉTABLIR ».

« METTRE SUR PIED » eût été meilleur. Mais là aussi, il aurait fallu créer un descripteur pour le GN2 : « force » seul cas d'espèce.

Dans ces deux exemples on a affaire à des traductions « **approximatives** » : la plus grande partie du sens est conservée.

– **Deuxième cause** de traduction non satisfaisante : le descripteur est **trop général**.

Exemple :

Very few if any among them HELD SLAVES.

Règle suivie : HOLD + GN2;

si GN2 = « HUMAIN » et GN1 = « HUMAIN » ⟶ « TENIR ».

Ici le descripteur « HUMAIN » sur le GN2 **slave** est trop vaste; il aurait fallu ajouter un trait comme « infériorité » ou « bien possédé » pour pouvoir traduire par « AVOIR » ou « POSSÉDER ». Mais là encore c'est le seul cas rencontré.

Ici la traduction est « impropre » : le sens source est perdu mais cela reste compréhensible.

– **Troisième cause** de traduction non satisfaisante : le **caractère limité** de nos investigations qui ne portent que sur la phrase et donc sur un environnement trop étroit.

Exemple :

Sometimes he **DREW** birds on his copy-book.

Règle : DRAW + GN2; si GN2 = « ANIMAL » ⟶ « ATTIRER » au lieu de « DESSINER ».

Il aurait fallu pouvoir détecter une information sur le circonstant et la mettre en relation avec les autres éléments. Il faudrait faire appel à des techniques d'IA pour formaliser un univers textuel plus vaste.

Ici la traduction est totalement **incorrecte** : perte de sens.

– **Dernière cause** : **l'ambiguïté** totale du verbe dans la phrase. Dans ce cas le système automatique sort deux traductions possibles et c'est au réviseur humain à choisir.

Exemple : Steve **DREW** a glass of water.

Il est impossible de choisir entre « TIRER » ou « DESSINER » car ici aussi le contexte est trop étroit; il faudrait prendre en compte les informations données antérieurement dans le récit.

Conclusion

À partir de cette étude, limitée alors à cinq verbes, on peut constater l'aide effective apportée par l'environnement syntactico-sémantique au transfert lexical en TA. Les premiers résultats obtenus par une recherche empirique peuvent donner l'impression de manquer de bases solides; c'est pourquoi il nous apparaît que tout progrès dans l'automatisation de la démarche de traduction exige des recherches plus poussées sur un contexte plus élargi, notamment en direction de l'IA.

Quant aux descripteurs sémantiques que nous avons élaborés jusqu'ici, ils sont appelés à évoluer en fonction de l'extension de notre étude à d'autres verbes. Certains descripteurs disparaîtront, d'autres fusionneront, d'autres seront créés et d'autres se confirmeront. À ce propos, par exemple, nous avons constaté que le descripteur « PORTEUR D'INFORMATION » déjà productif pour DRAW, HOLD et TURN, se révèle à nouveau utile pour certaines traductions de RUN et MOVE étudiés récemment.

Enfin, on peut dire que le fait **d'extraire les seuls éléments sémantiques indispensables** pour le choix de traductions dans nos types de textes, est préférable à l'application d'une théorie sémantique préétablie.

Annexe 1

1. The senate seat *held* by her husband Richard.
   occuper
2. He *held* a reception.
   organiser
3. The transplant surgeon *was held* the immediate cause of the death.
   tenir... pour
4. The century dinner *was held* at Eaton.
   avoir lieu
5. ...the man who *holds* the record
   détenir
6. To *hold* a herd of cattle on the range.
   garder
7. Rosemary's mother was *holding out* her long hand.
   tendre
8. Research workers *hold* an opposite opinion.
   soutenir
9. He *was holding* a carabin.
   tenir
10. Inventions are *held up* time and again.
    retarder
11. The incubator *holds* yoghourt jars.
    contenir

12. Joe *held up* the book again.
présenter
13. The technique *holds* great promise...
contenir
14. Party leaders *holding* shares in private firms...
détenir
15. A surplus of labor *holds back* the mechanization of agriculture.
freiner

Annexe 2

AFFECTIVITÉ	**FINANCE**
admiration	benefit
fury	dollar
glamor	money
hatred	salary
suspicion...	share...

PORTEUR D'INFORMATION	**PORTEUR DE VIRTUALITÉS**
data	future
idea	hope
letter	menace
page	possibility
problem...	promise...

Annexe 3

position	entité spatiale (draw, hold, lie)
position	lieu (hold, work)
position	opinion (hold)
position	socioprofessionnel abstrait (hold)
book	contenant (hold, turn)
book	objet (draw, hold, turn)
book	porteur d'information (draw, hold, turn)

Annexe 4

PHÉNOMÈNE	**SOCIAL**	**TEMPORAIRE**
census	debate	meeting
ceremony	dinner	service
congress	election	show
course	festival	trial...

ENTITÉ	**SPATIALE**	**et GÉOGRAPHIQUE**
area	hill	place
city	lake	pool
field	land	sea
glacier	park	valley...

3

Problèmes de désambiguïsation interactive et TAO personnelle

Hervé BLANCHON

GETA, Institut IMAG (UJF & CNRS), Grenoble, France

Résumé

Nous nous plaçons ici dans un contexte de recherche autour d'un système de *TAO personnelle*. L'utilisateur d'un tel système est un rédacteur unilingue qui souhaite pouvoir rédiger dans sa langue maternelle et être traduit pour un coût et dans un délai raisonnables. Pour la réalisation, on doit respecter trois contraintes très fortes. D'abord, la qualité de la traduction obtenue doit être suffisante pour ne pas devoir réviser (le rédacteur n'est pas un réviseur). Ensuite, on ne peut pas utiliser de base de connaissances pour obtenir automatiquement une compréhension explicite du document (il est impossible de connaître *a priori* le domaine sémantique de celui-ci). Enfin, on ne peut pas contraindre le rédacteur à utiliser un langage contrôlé, ni même un sous-langage déterminé (la rédaction doit être la plus libre possible).

Le respect des deux dernières contraintes ne permet pas la réalisation d'un système autonome de traduction avec lequel les analyses (syntaxique, logico-sémantique) du texte en langue source reflètent l'intention de l'auteur. Le système produira en effet, à certains stades du traitement, des représentations qui refléteront des ambiguïtés que les connaissances grammaticales « fortes » et les connaissances sémantiques « faibles » du système ne permettront pas de résoudre automatiquement. Or, si l'on veut obtenir une traduction de qualité, il faut obtenir une représentation désambiguïsée de ce que l'on veut traduire. La solution que nous envisageons au GETA est de réaliser, dans le cadre du projet LIDIA, la maquette d'un système de traduction qui utilise le dialogue avec le rédacteur pour obtenir une représentation désambiguïsée de ce que l'on veut

traduire. Dans un tel système, le dialogue occupe donc une place privilégiée puisque c'est lui qui va permettre d'obtenir une traduction de qualité.

Dans le présent article, nous décrirons comment, à partir d'un résultat d'analyse, nous percevons l'organisation des traitements qui nous permettront d'obtenir une représentation débarrassée de trois types d'ambiguïtés. Pour chacun de ces types, nous montrerons quel genre de dialogue peut proposer le système. Nous verrons alors que l'on peut adapter la forme du dialogue à l'utilisateur. Pour un utilisateur novice on peut envisager un dialogue très explicite. En effet, nous allons parfois surprendre le rédacteur en lui posant certaines questions. Il faut alors que l'explicitation du phénomène linguistique soit très précise et suffisamment simple pour être comprise sans avoir à faire appel à des connaissances linguistiques pointues. Ce type de dialogue doit aider le rédacteur à reconnaître les schémas ambigus pour éviter leur emploi ou bien être conscient qu'il devra répondre à des questions les concernant. Ensuite, avec une certaine habitude et une amélioration de ses compétences, il pourra choisir un mode de dialogue plus synthétique, qui rendra la tâche de désambiguïsation plus rapide.

Situation : l'analyseur de LIDIA

Les traitements interactifs dans le projet LIDIA se divisent en deux catégories : les traitements de standardisation et les traitements de clarification (Blanchon 1990, Boitet 1990). Lorsque les traitements de standardisation ont été effectués sur une unité de traduction, une première étape d'analyse produit une structure « concrète » qui est un arbre décoré « multiniveau » (Vauquois 1978). La structure est dite « concrète », car on y retrouve directement le texte (mot des feuilles). Elle est « multiniveau » car les nœuds portent des décorations complexes qui représentent divers niveaux d'interprétation : le niveau des classes syntaxiques et des classes syntagmatiques, le niveau des fonctions syntaxiques et enfin le niveau des relations logiques et sémantiques. Une première phase de désambiguïsation interactive s'applique à cette structure.

Pour préciser les choses, voici une brève description de ces divers niveaux d'interprétation qui permettra de mieux comprendre notre stratégie de désambiguïsation.

> Le niveau des classes syntaxiques (classes terminales telles que : « NOM COMMUN », « ADJECTIF », « PRÉPOSITION », etc.) et des classes syntagmatiques (classes non terminales telles que : « GROUPE NOMINAL », « PHRASE INFINITIVE », « PHRASE VERBALE », etc.) donne le parenthésage de la phrase en groupes syntagmatiques. Les feuilles de l'arbre portent les classes syntaxiques et les nœuds non terminaux portent les classes syntagmatiques.
>
> Le niveau des fonctions syntaxiques donne le rôle syntaxique de chaque nœud dans le groupe auquel il est attaché directement (par exemple : « GOUVERNEUR », « RÉGISSEUR », « SUJET », etc.). Avec ce niveau, on peut très simplement transformer l'arborescence en une arborescence de « dépendance syntaxique ».

Le niveau des relations logiques et sémantiques peut être considéré comme celui de la construction logique de la phrase, en mettant en évidence les places des arguments attachés aux unités lexicales prédicatives et donnant l'interprétation sémantique des compléments et circonstants (par exemple : « DÉTERMINATION », « CIRCONSTANCIEL INSTRUMENTAL », « PREMIER ARGUMENT D'UNE UNITÉ LEXICALE PRÉDICATIVE », etc.).

L'arbre d'analyse de la phrase: « L'élève calcule cette intégrale par la méthode des résidus » tel qu'il est classiquement calculé au GETA, est fourni pour clarification.

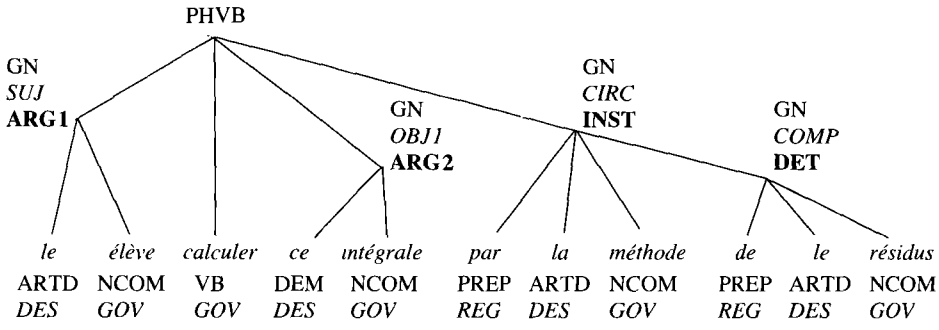


Figure 1 : Un arbre d'analyse multiniveau.

légende : NIVEAU 1, *NIVEAU 2*, **NIVEAU 3**

L'analyseur utilisé dans le cadre du projet LIDIA est un analyseur non déterministe « multiple » qui, pour une phrase donnée, fournit toutes les analyses vérifiant le modèle syntagmatique, syntaxique et logico-sémantique de l'analyseur. La figure 2 montre la forme d'une analyse produite pour une unité textuelle selon le modèle non déterministe. Pour la première phrase, il y a différentes analyses : le nœud père « UL(ULFRA),NUMFRA(1) » possède plusieurs fils UL(ULSOL) et, sous chaque nœud UL(ULSOL), on dispose d'un arbre d'analyse licite de la phrase.

On impose aussi, par rapport au modèle habituel, une contrainte supplémentaire à l'analyseur : le mot qui appartient à la décoration d'une feuille de l'arbre (*en italique dans la figure 1*) doit être une occurrence de la phrase analysée, et la projection du mot des feuilles de l'arbre selon cette composante mot de la décoration fournit la phrase analysée. Le respect de cette contrainte nous permet d'avoir un processus de génération des questions plus simple à mettre en œuvre et donc moins coûteux en logiciel et en temps. Le processus de génération du dialogue étant exécuté par la station de rédaction, on perçoit l'intérêt qu'il y a à minimiser l'importance des tâches à réaliser pour que leur exécution soit la plus transparente possible pour l'utilisateur. Il ne faut pas que

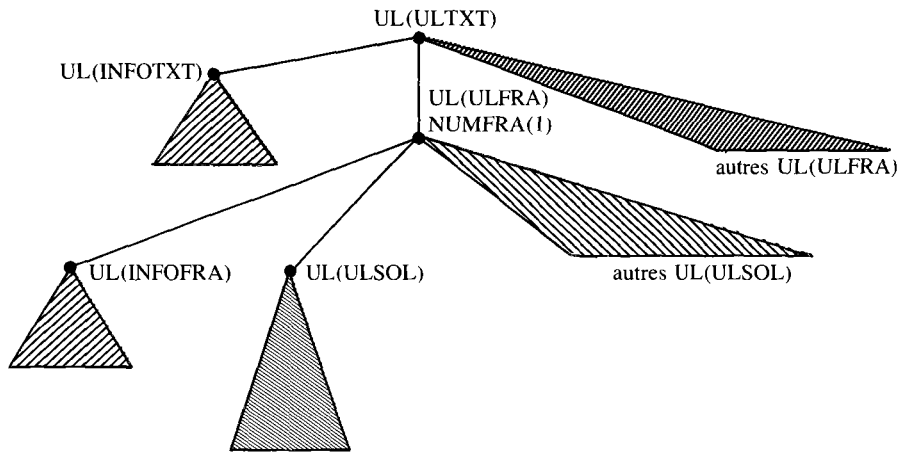


Figure 2 : Représentation d'une analyse non déterministe.

des tâches de fond trop importantes ralentissent le travail du rédacteur de façon significative.

Pour procéder à la désambiguïsation interactive en utilisant le résultat d'analyse comme support pour la préparation des questions, nous avons choisi d'utiliser trois propriétés qui permettent, pour une phrase, de différencier les divers arbres obtenus. Ces trois propriétés sont : l'étiquetage en classe syntaxique du mot des feuilles, l'étiquetage en fonction syntaxique (qui induit parfois des géométries différentes) et l'étiquetage en relation logique et sémantique des nœuds non terminaux de l'arbre. On va maintenant donner une définition de ces trois propriétés agrémentée d'exemples.

Propriétés choisies de la structure produite par l'analyseur

– La première propriété qui différencie les arbres d'analyse licite d'une phrase est l'étiquetage en classe syntaxique du mot des feuilles (désigné dans la suite par étiquetage en classe syntaxique) qui est, rappelons-le, le premier niveau d'interprétation de la structure profonde.

Le schéma suivant (figure 3) montre les différences entre les deux arbres produits par l'analyse de la phrase « Devant cette somme, il hésite » dans laquelle *devant* peut être interprété comme gérondif du verbe devoir ou bien comme préposition.

Comme on peut le voir, les arbres ont une géométrie différente (étiquetage en fonction syntaxique différent), mais ce qui nous intéresse ici c'est seulement la différence au niveau de l'étiquetage en classe syntaxique de l'occurrence *devant*.

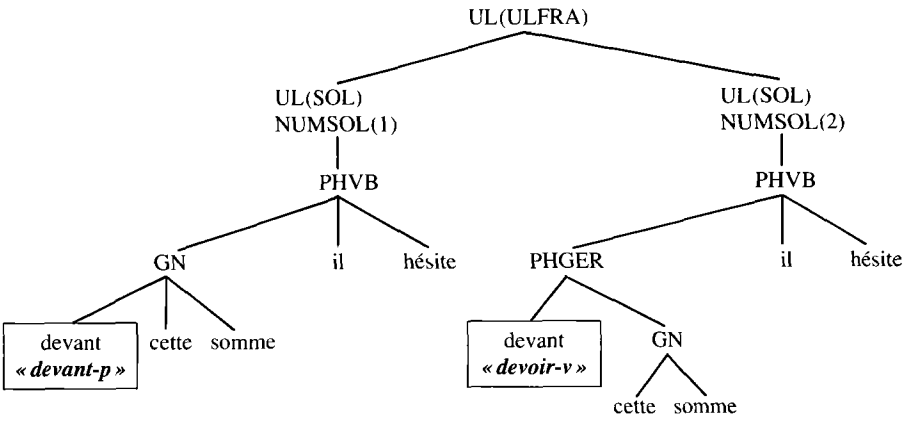


Figure 3 : Résultat d'analyse de la phrase « Devant cette somme, il hésite. » On observe un problème au niveau du choix de la classe syntaxique pour l'occurrence **devant**.

– La seconde propriété qui différencie les arbres d'analyse licite d'une phrase est la géométrie et l'étiquetage en fonction syntaxique des nœuds non terminaux (désigné par la suite par géométrie et étiquetage en fonction syntaxique). Les arbres ont maintenant tous le même étiquetage en classe syntaxique. Cette propriété correspond au deuxième niveau d'interprétation de la structure profonde.

Le schéma suivant (figure 4) montre les différences entre les deux arbres produits par l'analyse de la phrase « L'évolution de la structure du réseau et des investissements est étudiée » où l'on peut comprendre que l'on étudie l'évolution de la structure du réseau **et** l'évolution des investissements (NUMSOL(1)) ou bien que l'on étudie l'évolution de la structure du réseau **et** l'évolution de la structure des investissements (NUMSOL(2)).

– La troisième propriété qui différencie les arbres d'analyse licite d'une phrase est l'étiquetage en relation logique et relation sémantique des nœuds non terminaux (désigné par la suite par étiquetage en relation logico-sémantique). Elle correspond au troisième niveau d'interprétation de la structure profonde.

Le schéma suivant (figure 5) montre les différences entre les deux arbres produits par l'analyse de la phrase « Je vous parle de la tour Eiffel » où on peut comprendre que l'on nous parle depuis la tour Eiffel (NUMSOL(2)) ou bien que l'on nous parle au sujet de la tour Eiffel (NUMSOL(1)).

Après avoir décrit les propriétés qui nous intéressent, montrons comment, en exploitant leur présence dans le résultat d'analyse d'une phrase, nous pouvons résoudre certains types d'ambiguïtés.

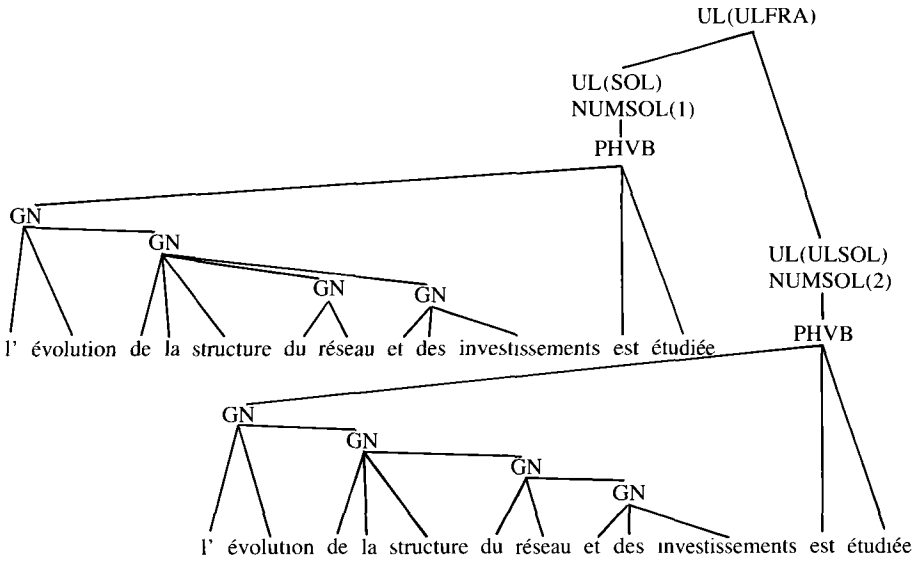


Figure 4 : Résultat d'analyse de la phrase « L'évolution de la structure du réseau et des investissements est étudiée. » On observe un problème au niveau du choix de la géométrie des arbres.

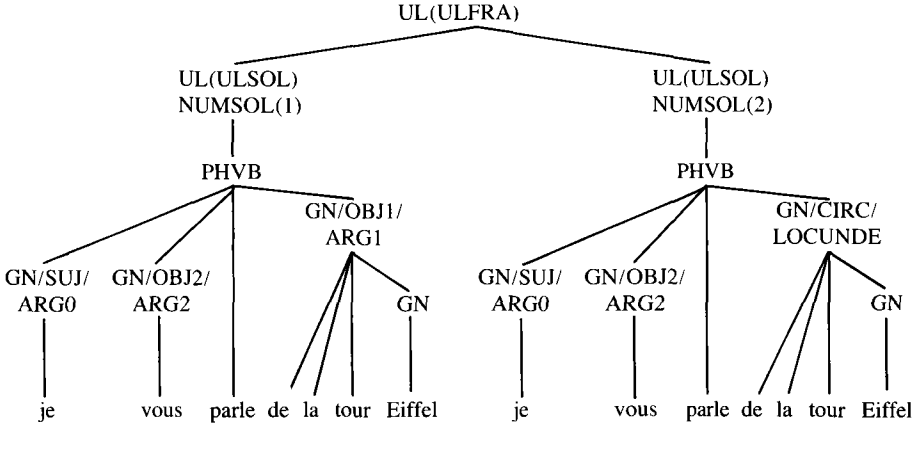


Figure 5 : Résultats d'analyse de la phrase « Je vous parle de la tour Eiffel. » On observe un problème au niveau du choix de la fonction du groupe nominal **de la tour Eiffel**.

Intérêt des propriétés choisies

Relation entre propriété et type d'ambiguïté

Si l'analyse indique un problème d'étiquetage en classe syntaxique, c'est l'ambiguïté de classe syntaxique que nous avons à résoudre. Cela signifie que l'analyseur, pour une occurrence donnée, n'a pas su déterminer quelle est sa classe syntaxique. Par exemple, dans la phrase « Le vétérinaire prend la patte de l'oiseau et **la** bague », on ne sait pas, *a priori*, si l'occurrence *bague* joue le rôle d'un verbe (**la** est alors pronom) ou d'un substantif (**la** est alors article).

Par contre, si l'analyse indique un problème de géométrie et d'étiquetage en fonction syntaxique, c'est l'ambiguïté de structure de constituants grammaticaux que nous devons résoudre. Cette ambiguïté de constituants grammaticaux est essentiellement liée à deux problèmes : la coordination et la subordination. La phrase « Il prend des classeurs et des cahiers noirs » présente un problème de géométrie pure, puisque l'on ne sait pas s'il faut comprendre que seuls les cahiers sont noirs ou bien que les classeurs sont noirs et les cahiers aussi. Pour la subordination, on peut citer la phrase : « Il est inscrit dans une école de cuisine lyonnaise » où on ne sait pas s'il s'agit d'une *école située à Lyon où on apprend la cuisine* ou bien d'une *école sise on ne sait où, où on apprend la cuisine de Lyon.*

Enfin, lorsque l'analyse indique un problème d'étiquetage en relation logico-sémantique, on a à résoudre deux types de problèmes : un problème de place des arguments attachés aux unités lexicales prédicatives (ex. : « Quel hôtel gère ce bureau? »), ou bien un problème d'interprétation sémantique des compléments et circonstants. Comme exemple de problème d'interprétation, on peut citer la phrase « Je vous parle de la tour Eiffel » où *de la tour Eiffel* peut être considéré comme circonstant de localisation (depuis la tour Eiffel), ou comme premier argument et premier objet du verbe parler (au sujet de la tour Eiffel).

Combinaisons possibles des propriétés

Bien sûr, une phrase peut donner lieu à plusieurs ambiguïtés, de même type ou de types différents. Tous les croisements sont effectivement possibles, comme le montrent les exemples suivants.

- Différents étiquetages en classes syntaxiques du mot des feuilles :
  - avec différentes géométries et étiquetages en fonctions syntaxiques et avec différents étiquetages en relations logico-sémantiques :

    Il nous parle de l'école de cuisine lyonnaise, prend le livre et le lit.

    On voudrait savoir comment interpréter
    « de », « de » : « depuis », « au sujet de »
    « école de cuisine lyonnaise » : « école lyonnaise », « cuisine lyonnaise »
    « le lit » : « pronom + verbe », « article + substantif »

- avec différentes géométries et étiquetages en fonctions syntaxiques et un seul étiquetage en relations logico-sémantiques :

  Il regarde le livre de l'école de cuisine lyonnaise et le lit.

- avec une seule géométrie et étiquetage en fonctions syntaxiques et avec différents étiquetages en relations logico-sémantiques :

  Il observe la photo de Paul et la classe.

  On voudrait savoir comment interpréter
  « de » : « prise par Paul », « représentant Paul », « appartenant à Paul »
  « la classe » : « pronom + verbe », « article + substantif »

- avec une seule géométrie et étiquetage en fonctions syntaxiques et un seul étiquetage en relations logico-sémantiques :

  Le vétérinaire prend la patte de l'oiseau et la bague.

  On voudrait savoir comment interpréter
  « la bague » : « pronom + verbe », « article + substantif »

- Différentes géométries et étiquetages en fonctions syntaxiques :
  - avec différents étiquetages en relations logico-sémantiques :

    Il parle de l'école de cuisine lyonnaise.

    On voudrait savoir comment interpréter
    « de » : « depuis », « au sujet de »
    « école de cuisine lyonnaise » : « école lyonnaise », « cuisine lyonnaise »

  - avec un seul étiquetage en relations logico-sémantiques :

    Il est inscrit dans une école de cuisine lyonnaise.

- Différents étiquetages en relations logico-sémantiques seulement :

  Quel hôtel gère ce bureau?

  On voudrait savoir comment interpréter les arguments du verbe
  « un hôtel gère-t-il le bureau? » ou « le bureau gère-t-il un hôtel? »

Définition d'un ordre d'utilisation des propriétés

Comme plusieurs propriétés peuvent apparaître dans un résultat d'analyse, nous allons exploiter, si elles se présentent, les propriétés dans l'ordre dans lequel nous les avons définies. C'est-à-dire l'étiquetage en classes syntaxiques, puis la géométrie et l'étiquetage en fonctions syntaxiques et enfin l'étiquetage en relations logico-sémantiques. La figure 6 donne une idée de la façon dont nous allons procéder. Le but du processus que nous allons décrire est la production d'une série de questions organisées hiérarchiquement qui permettent d'obtenir, en partant d'une liste d'analyses licites, l'analyse qui correspond aux intentions du rédacteur. Cet ordre d'exploitation des propriétés pourra, si nécessaire, être remis en question car, comme le montre la figure 6, l'organisation des traitements est fortement modulaire.

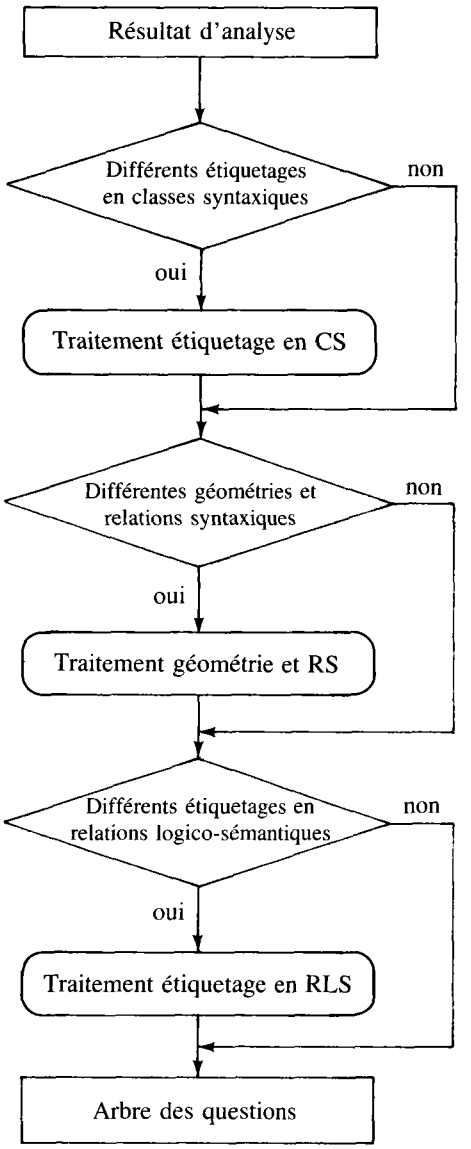


Figure 6 : Schéma du processus d'exploitation de l'arbre d'analyse.

On est libre d'examiner la présence de chacune des propriétés dans n'importe quel ordre. Mais il nous semble que l'organisation proposée a quelques fondements psychologiques et ergonomiques.

Pour le point de vue psychologique, lorsque l'on se place dans la situation de levée interactive d'ambiguïté, le rédacteur sera amené à réexaminer son texte sans se servir du contexte, comme le fait le système. Supposons par exemple qu'il soit occupé à rédiger un document qui parle du classement d'un certain nombre de clichés représentant des acteurs célèbres ou en train de le devenir et que dans un tel contexte il écrive la phrase « Il examine la photo de Pierre Latouche et la classe ». Dans ce texte, cette phrase ne pose aucun problème d'interprétation. Mais, sortie de tout contexte, elle est ambiguë, puisqu'on peut comprendre qu'il classe des photos ou qu'il examine une photo et une classe (un groupe d'élèves par exemple).

Hors contexte, la fabrication du sens (par un humain) semble procéder par restriction des classes syntaxiques des termes, puis des géométries et relations de coordination et de subordination qu'ils établissent ou non avec d'autres termes et enfin, par la qualité de ces relations. D'un point de vue psychologique, il semble donc convenable de procéder de manière analogue.

Le point de vue ergonomique renforce cette idée : il est préférable de réussir rapidement la désambiguïsation en posant les questions par ordre d'immédiateté de leurs réponses. C'est dire que la sélection de la bonne analyse doit procéder par « filtrages » successifs en utilisant d'abord des critères simples et des questions courtes, et ensuite seulement des critères plus fins et donc des questions plus complexes.

Ainsi, pour trouver la bonne interprétation de la phrase « Il atteint la grange et la ferme », il est plus naturel de demander si l'occurrence *ferme* doit être comprise comme le verbe *fermer* ou comme substantif que de demander si *et la ferme* doit être compris comme un groupe nominal coordonné au verbe *atteindre* ou bien comme une phrase verbale coordonnée au verbe *atteindre*. Il est donc normal de commencer par chercher le bon étiquetage en classes syntaxiques puis la bonne géométrie et le bon étiquetage en fonctions syntaxiques, et enfin le bon étiquetage en relations logico-sémantiques.

Utilisation des propriétés choisies en désambiguïsation

Généralités

Le système partitionnera donc l'ensemble des solutions en trois étapes successives, d'abord selon l'étiquetage en classes syntaxiques, puis selon la géométrie et l'étiquetage en relations syntaxiques et enfin, selon l'étiquetage en relations logico-sémantiques. On aboutira donc à un arbre des questions dont les niveaux successifs correspondront à cette organisation tripartite.

La figure 7 donne un exemple de découpage en classes et montre quel est l'arbre de décision que l'on peut construire pour hiérarchiser les questions.

Cette figure montre que les analyses A 1, A 2 et A 3 ont le même étiquetage en classes syntaxiques et que A 1 et A 2, tout en ayant le même étiquetage en fonctions syntaxiques, ont des étiquetages en relations logico-sémantiques différents; A 3, lui, a un autre étiquetage en fonctions syntaxiques. De même les analyses A 4 et A 5 ont le même étiquetage en classes syntaxiques, mais des étiquetages différents en relations syntaxiques.

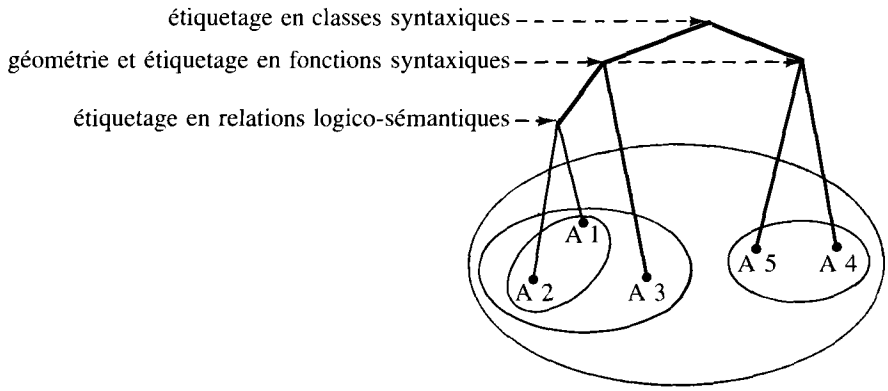


Figure 7 : Arbre de décision permettant de sélectionner la bonne analyse en fonction de la partition des analyses.

Exemple

Examinons la phrase « Le professeur de l'école de cuisine lyonnaise parle de la Tour d'Argent, prend le menu et le lit ». Cette phrase a seulement la prétention d'être un exemple démonstratif.

- Sans factorisation, l'analyse de cette phrase fournit huit solutions.
- On a schématiquement l'arbre de questions suivant :

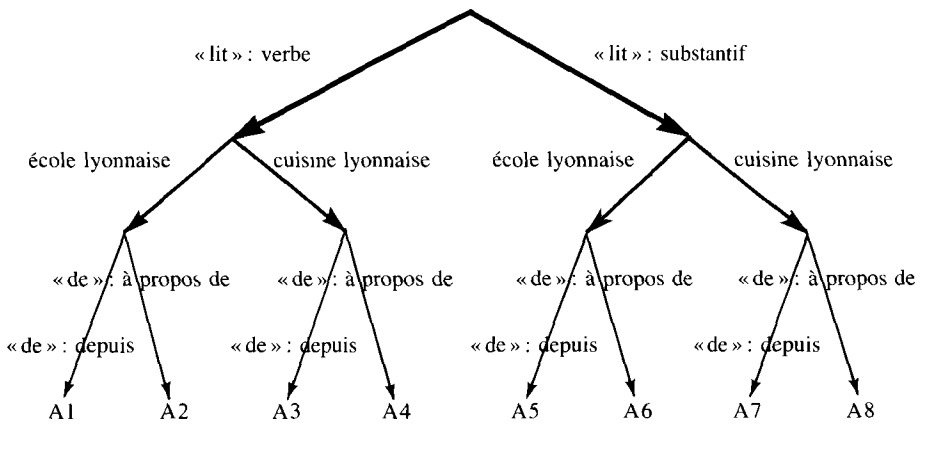


Figure 8

– Pour les questions on a les dialogues suivants :

– Pour la résolution des ambiguïtés de classes, on propose trois types de dialogue :

1) Une présentation grammaticale explicite du problème, pour se familiariser avec le système.

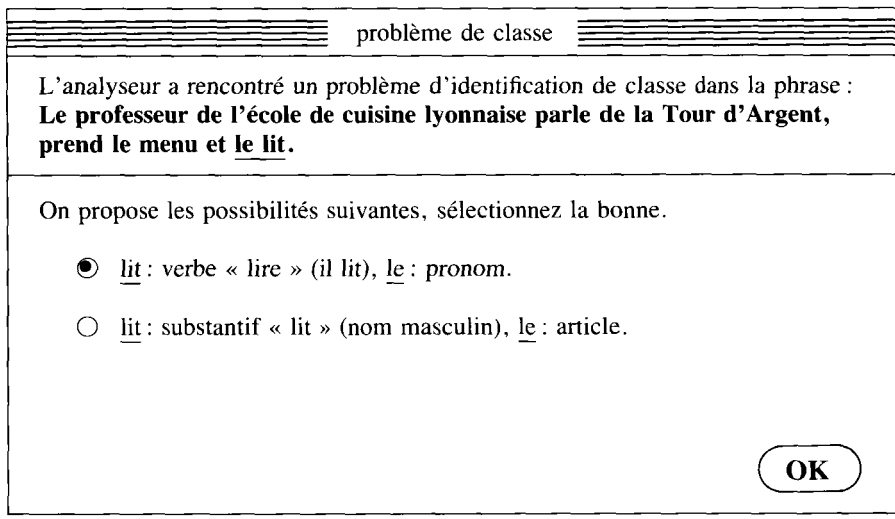


Figure 9

2) Un paraphrasage de la partie de la phrase qui pose problème, l'explication grammaticale est implicite.

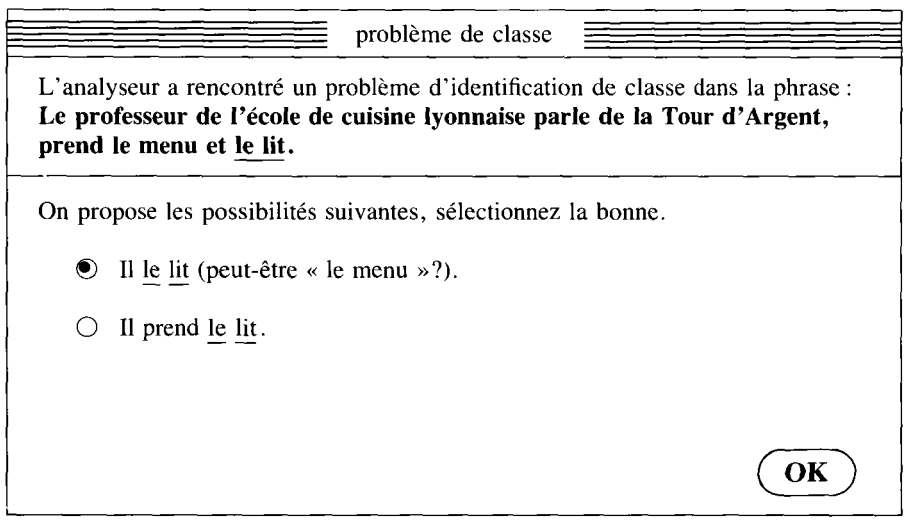


Figure 10

3) Une projection des classes syntaxiques de la partie de la phrase qui pose problème, destinée à un utilisateur qui a voulu se familiariser avec une présentation grammaticale des problèmes.

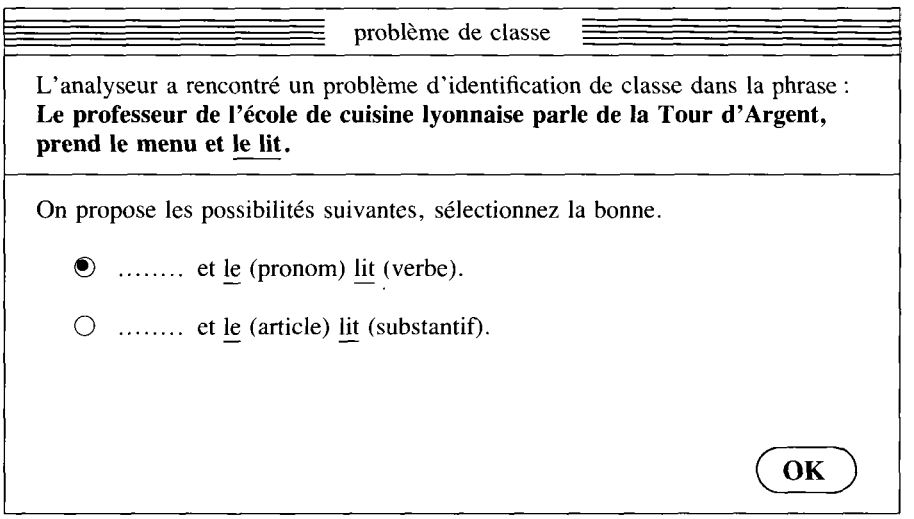


Figure 11

– Pour la résolution des ambiguïtés de constituants grammaticaux, on peut proposer des présentations de type paraphrastique qui tiennent plus ou moins compte du contexte

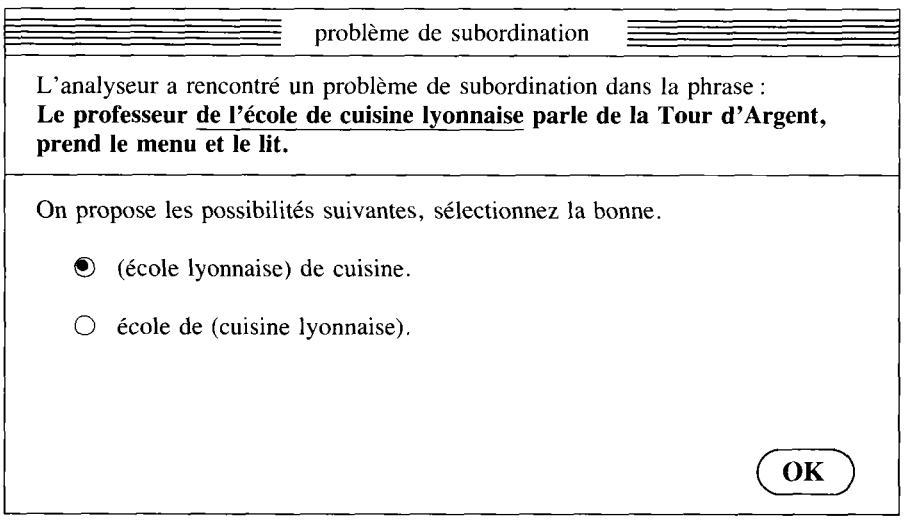


Figure 12

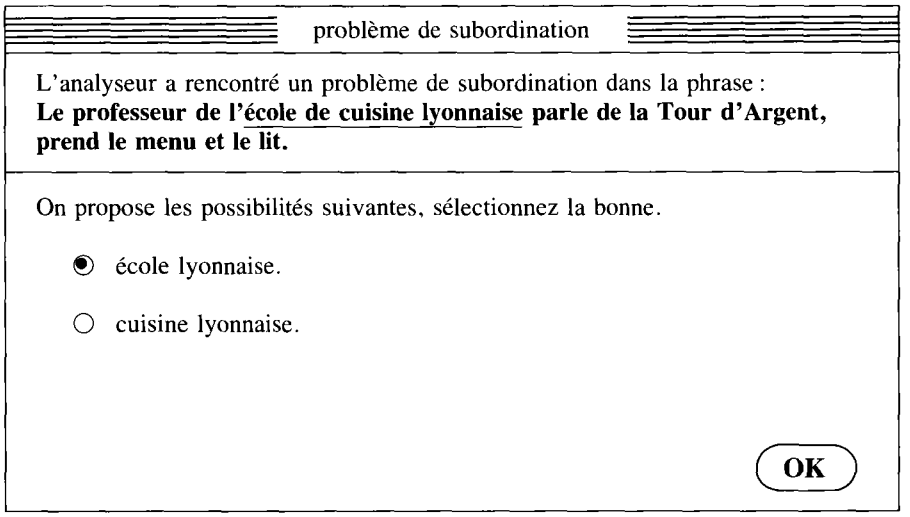


Figure 13

ou une présentation plus compacte qui exige peut-être un peu plus d'expérience du système.

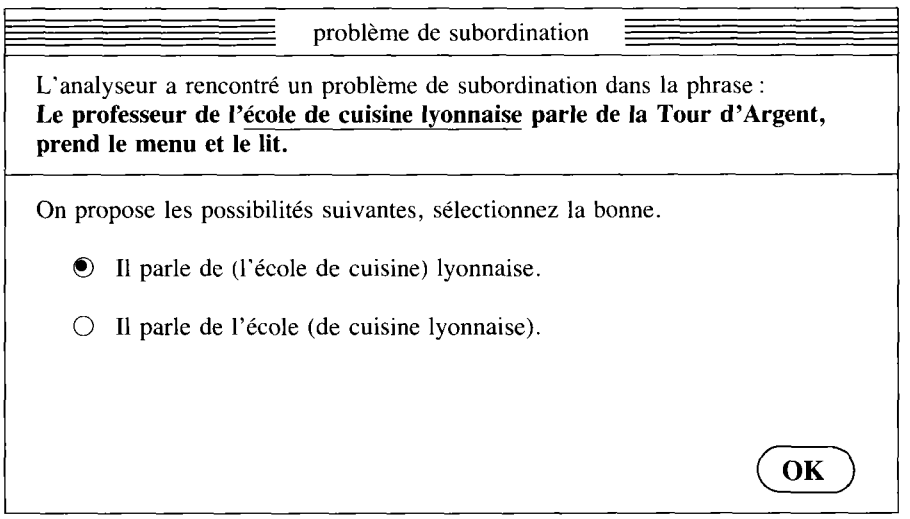


Figure 14

- Pour la résolution des ambiguïtés de relation logico-sémantique, on peut ici proposer des niveaux très différents de dialogue :
  1) Un niveau où on remplace l'occurrence qui permet de mettre en évidence la place des arguments ou d'introduire les compléments ou les circonstants. Ici,

il s'agit de remplacer *de* qui permet d'introduire, soit l'argument 1 du verbe parler soit un circonstant de localisation.

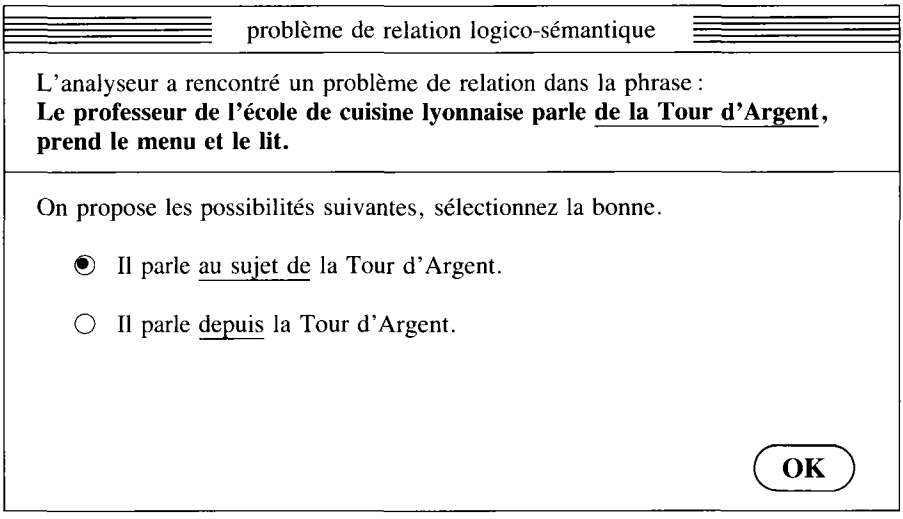


Figure 15

2) Un niveau qui précise en plus l'étiquetage de l'occurrence *de* qui permet d'apprendre les représentations du système.

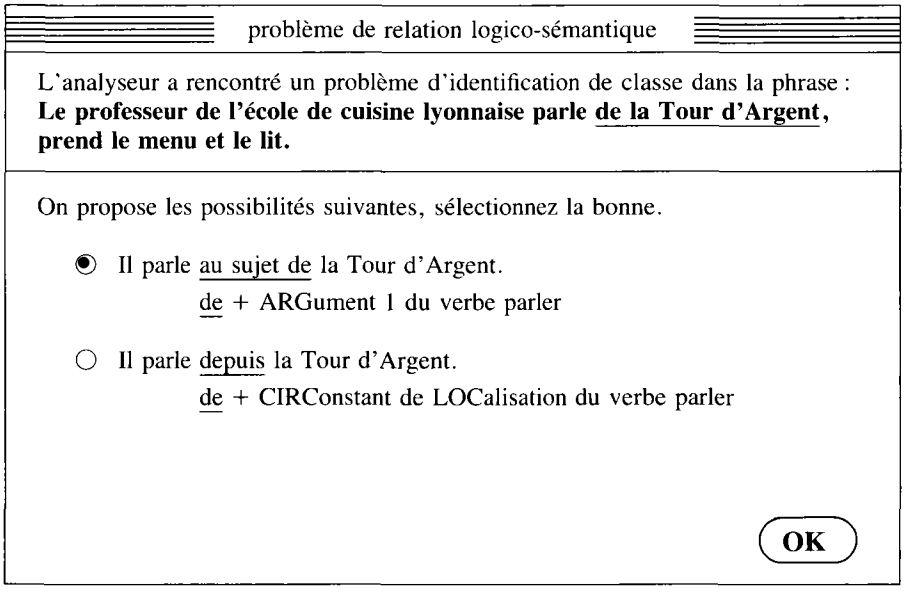


Figure 16

3) Un niveau pour un utilisateur spécialiste.

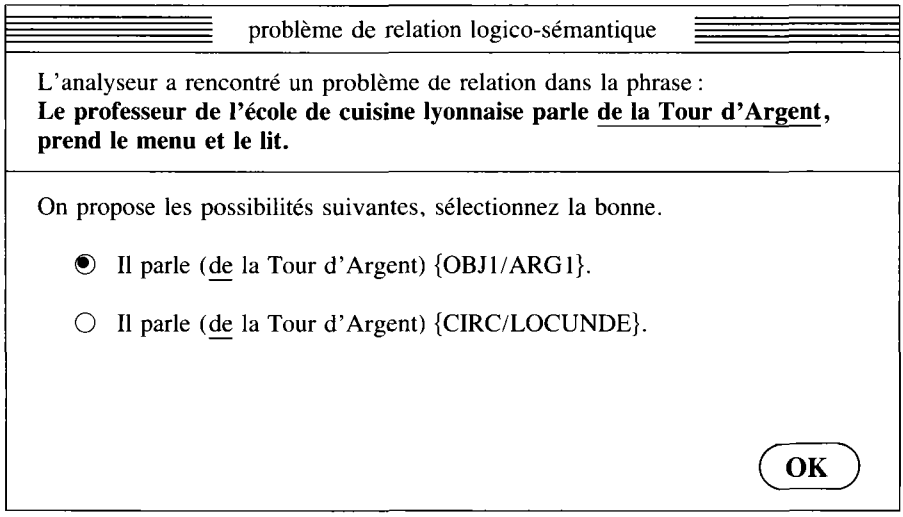


Figure 17

Conclusion

Nous avons défini trois propriétés de la structure arborescente concrète, mais profonde (en fait, multiniveau) produite par un analyseur non déterministe (dans le formalisme du GETA). Nous avons vu comment exploiter ces propriétés dans le cadre d'un système de « TAO personnelle » pour produire un dialogue de désambiguïsation qui soit réglable en fonction des compétences de l'utilisateur. Les ambiguïtés que l'on résout sont, on le rappelle, les ambiguïtés de classe syntaxique, les ambiguïtés de coordination et de subordination, et enfin les ambiguïtés de relation logique et sémantique. Mais, pour obtenir une représentation totalement désambiguïsée il restera d'autres problèmes à régler.

C'est ainsi qu'il faut encore résoudre les problèmes de polysémie (pure et induite par la langue cible), les problèmes de référence pronominale ou adjectivale et les problèmes d'ellipse, soit pure (toutes les fois qu'un terme de phrase ou de groupe n'est pas repris, souvent dans la coordination ou la comparaison), soit dans les termes composés (lorsque l'on ne reprend pas complètement un terme composé pour abréger ou alléger le texte, le terme est remplacé par son gouverneur).

Les phases de désambiguïsation qui traitent de ces problèmes sont réalisées soit en amont soit en aval des traitements que nous avons décrits ici. Pour information, on peut dire que pour les polysémies, la notation du problème sera inscrite dans l'arbre dans l'étiquette de sens associée à chaque occurrence, et les informations nécessaires à la désambiguïsation seront fournies par la base lexicale. Pour les problèmes de référence pronominale, il faudra conserver une liste de toutes les occurrences qui

peuvent être reprises en associant à chacune d'elles des informations grammaticales (genre, nombre...) et des informations lexicales provenant du dictionnaire (les rections syntaxiques et traits sémantiques). Ces informations permettront de retrouver les occurrences pertinentes candidates à la reprise. Quant au problème de l'ellipse dans les termes composés, une liste des termes composés déjà rencontrés ou des informations à propos du domaine du document permettront de trouver une liste des candidats possibles.

Bibliographie

BLANCHON H. (1990b) « HyperAriane : architecture de la partie *micro* d'un système de TAO personnelle », *Mémoire de DEA*, juin 1990, 63 p.

BLANCHON H. (1990a) « LIDIA-1 : Un prototype de TAO personnelle pour rédacteur unilingue », *Proc. Avignon-90, conférence spécialisée Le Traitement des langues naturelles et ses applications*, 28 mai–1er juin, 1990, p. 51-60.

BOITET Ch. (1990b) « Vers la TAO personnelle : le projet LIDIA du GETA », à paraître dans *la tribune sur les industries de la langue*, 12 p.

BOITET Ch. (1990a) « Towards Personal MT : On Some Aspects of the LIDIA Project », *Proc. COLING-90*, Helsinki, 20-25 août 1990, 10 p.

BOITET Ch. (1989b) « Speech Synthesis and Dialog Based Machine Translation », *Proc. ATR Symp. on Basic Research for Telephone Interpretation*, Kyoto, Dec. 1989, 6-5-1–6-5-22.

BOITET Ch. (1989a) « Motivation and Architecture of the LIDIA Project », *Proc. MTS-89*, Munich, 16-18 août 1989, 5 p.

BOITET Ch. et BLANC E. (éds.) (1990) *DBMT-90 : Post-COLING Seminar on Dialogue-Based MT (Machine Translation of/with Dialogues)*, Le Sappey (France), 26-28 août 1990, 328 p.

BROWN R. D. et NIREMBURG S. (1990) « Human-Computer Interaction for Semantic Disambiguation », *Proc. COLING-90*, Helsinki, 20-25 août 1990, 6 p.

HUTCHINS W. J. (1989) « Out of the Shadows, A Retrospect of Machine Translation in the Eighties », Presented at *Computer and Translation 89*, Tbilissi, Georgia, novembre-décembre 1989, 13 p.

LEPAGE Y. (1988) « Ambiguities and Second Generation MT Systems », *First European Conference on Information Technology for Organisational Systems*, Athens, 16-20 mai 1988, 6 p.

MARUYAMA H. et WATANABE H., OGINO S., (1989) « An Interactive Japanese Parser for Machine Translation », *Proc. COLING-90*, Helsinki, 2-25 août 1990, 7 p.

PECKHAM J. (ed) (1989) « Recent Developments and Applications of Natural Language Processing », Unicom Seminars Lilited, dans la collection *Applied Information Technology Reports*, 300 p.

TOMITA M. (1985) *Feasibility Study of Personal/Interactive Machine Translation System*, Center for Machine Translation, Carnegie Mellon University, mai 1985, 13 p.

TOMITA M. (1984) « Disambiguating Grammatically Ambiguous Sentences by Asking », *Proc. COLING-84*, p. 476-480.

VAUQUOIS B. (1978) « Description de la structure intermédiaire », Communication présentée au *colloque de Luxembourg*, 17 et 18 avril 1978, 25 p.

WHITELOCK P. J., WOOD M. M., CHANDLER B. J., HOLDEN N., HORSFALL H. J. (1986). « Strategies for Interactive Machine Translation : the Experience and Implications of the UMIST Japanese Project », *Proc. COLING-86*, p. 329-334.

ZAJAC R. (1986) *Étude des possibilités d'interaction homme-machine dans un processus de traduction automatique*, Thèse, INPG, Grenoble, juillet 1986, 259 p.

ZAJAC R. (1988) « Interactive Translation : a New Approach », *Report GETA & ATR*, Osaka, mai 1988, and *Proc. COLING-88*, Budapest, août 1988.

4

Dictionnaires électroniques dans un environnement traductionnel avancé

Ulrich HEID

Université de Stuttgart, Stuttgart, Allemagne

Avec la croissance du marché de la documentation, les outils informatiques deviennent de plus en plus importants; d'autre part, il subsiste, paraît-il, un écart considérable entre la performance des outils existants et les exigences des utilisateurs potentiels. Devant une telle situation, l'investissement dans les outils devrait s'accompagner de réflexions sur la compatibilité des systèmes avancés d'aujourd'hui avec les exigences du domaine, telles qu'elles se dessinent à travers les projets de recherche en cours.

Le domaine des ressources lexicales et terminologiques a déjà sérieusement entamé la discussion sur les questions de compatibilité et de réutilisation. Dans ce domaine, il existe déjà un certain nombre d'outils, pour la plupart interactifs, dont les chances d'intégration avec des applications avancées, telles que la génération automatique de textes, la traduction automatique, etc., soulèvent des doutes sérieux. Il faut alors s'interroger sur les caractéristiques que des ressources lexicales plus flexibles devraient avoir.

L'exposé part de ce scénario et donne une description rapide d'une base de données lexicale et terminologique conçue récemment en coopération avec l'entreprise Krupp Industrietechnik (Duisburg) qui constitue l'élément central d'une station de travail intégrée. À partir de travaux faits dans Polygloss, un projet de génération multilingue et d'EUROTRA-7, une étude de faisabilité sur la réutilisation de ressources lexicales, il faudra montrer les conditions nécessaires à une intégration d'outils interactifs et automatiques. L'accent sera mis sur les descriptions linguistiques nécessaires à une telle flexibilité.

5

Le module lexical dans un système de traduction automatique*

Liesbeth DEGAND, Guy EVERAERT, Jacques LEROT

Projet GENESE, Institut de linguistique, Louvain-la-Neuve, Belgique

Introduction

La génération et la traduction automatique de textes sont des procédés complexes dont la réalisation implique une architecture modulaire (modules syntaxique, discursif, sémantique et lexical). Ces modules sont en interaction permanente.

Dans le présent article, nous nous proposons de présenter le module lexical tel que nous l'envisageons dans notre système de génération automatique de textes.

Afin d'appréhender le lexique dans toute sa complexité, nous posons à l'intérieur du module lexical l'existence de trois niveaux :

- le modèle du monde (*world modelling structure*),
- le lexique conceptuel (*concept lexicon*),
- le lexique spécifique à chaque langue particulière (*language specific lexicon*).

Nous allons voir comment ces trois niveaux agissent les uns sur les autres pour fournir les lexèmes adéquats dans le processus de la traduction.

* La présente recherche, financée par la Communauté française de Belgique sous la convention n° 89-94/137, se fait dans le cadre d'un projet consacré à la génération de phrases françaises et néerlandaises à partir de représentations sémantiques. Il s'inscrit dans les efforts internationaux visant à développer un système de traduction multilingue (projet GENESE).

Le modèle du monde

Toute assertion fait appel aux connaissances générales appelées aussi connaissances du monde. Le *modèle du monde* est la composante qui représente ces connaissances générales[1].

Cette composante est représentée sous forme d'une hiérarchie où les éléments sont reliés entre eux de façon superordonnée et sousordonnée (relation qui sont exprimées par des arcs *une-sorte-de* (AKO — a kind of —)[2]. Le schéma suivant représente une telle organisation.

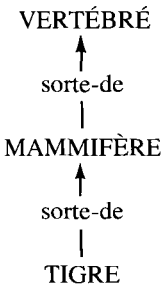


Figure 1

Cette représentation ne fournit qu'une information partielle à propos du monde. On décrira davantage les nœuds de la hiérarchie (appelés *types*) en spécifiant leurs propriétés typiques. On associe ainsi à chaque type un ensemble de paires attribut-valeur appelés *traits typés* (*feature types*) (Aït-Kaci 1986), (Emele, Zajac 1990). On représentera une telle paire de la manière suivante :

TYPE [attribut : valeur]

1. En réalité nous n'avons jamais besoin de toutes les connaissances du monde pour interpréter un fait, mais d'un sous-ensemble précis des connaissances générales (un sous-monde). Le modèle du monde est donc en fait un modèle du sous-monde (Nirenburg, Raskin 1987).
2. Pour une description plus complète des arcs, *cf.* Thayse 1988, chap.3.

Par exemple, le type « TIGRE » sera décrit à l'aide d'un ensemble de paires attribut-valeur tels que [possède : griffes], [est-une : espèce-en-danger], [sorte-de : mammifère], etc. On obtient le schéma suivant :

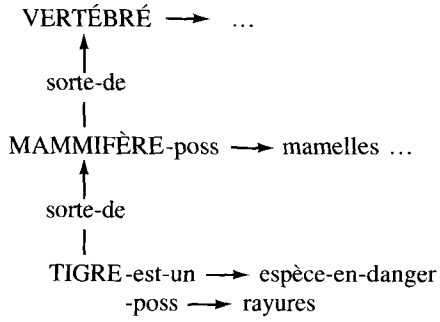


Figure 2

Ce système définit des types capables de représenter n'importe quelle réalité. Il permet donc d'intégrer des éléments mythiques imaginaires tels que, par exemple, les licornes, les griffons, etc.

Il y a plusieurs façons d'envisager ces hiérarchies qui varient selon les auteurs. Par exemple, pour Nirenburg et Carlson (1990), la construction se subdivise en deux sous-hiérarchies : ÉVÉNEMENT et OBJET.

Le type OBJET introduit les concepts OBJET-PHYSIQUE, OBJET-MENTAL et OBJET-SOCIAL. Les objets physiques peuvent à leur tour être classés en ENTITÉ SÉPARABLE, MATÉRIEL et LIEU. Parmi les objets mentaux, on reconnaît les OBJETs-ABSTRAITs et les OBJETs-REPRÉSENTATIFs (comprenant les OBJETs-MATHÉMATIQUEs, les OBJETs-LANGAGIERs, les ICÔNEs et les OBJETs-PICTURAUX). Finalement, les objets sociaux sont des ENTITÉs-GÉOPOLITIQUEs et des ORGANISATIONs.

Le type ÉVÉNEMENT est également subdivisible en un monde physique, mental et social. Les ÉVÉNEMENTs-PHYSIQUEs contiennent le CHANGEMENT-DE-LIEU et l'APPLICATION-DE-FORCE, les événements mentaux sont subdivisés en ÉVÉNEMENTs-COGNITIFs et ÉVÉNEMENTs-PERCEPTUELs, le monde social contient quant à lui des ÉVÉNEMENTs-COMMUNICATIFs.

Le sommet de la hiérarchie est représenté de la façon suivante (d'après Carlson L., Nirenburg S. 1990) :

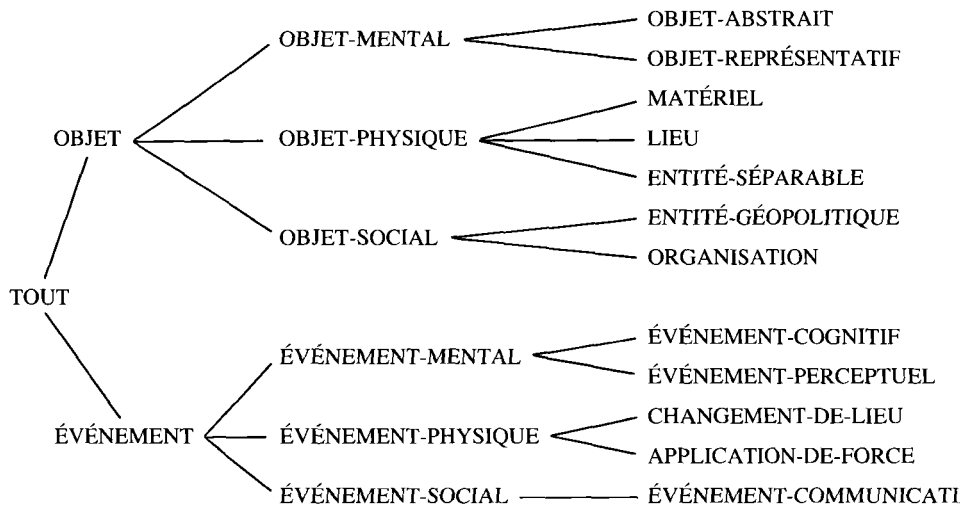


Figure 3

En résumé, nous pouvons dire que le modèle du monde est en quelque sorte un réservoir d'informations pertinentes qui sont à la base du processus de conceptualisation que nous décrirons dans la section suivante.

Il est important de noter que le modèle du monde est dynamique et qu'il peut être modifié et adapté en fonction des besoins de l'utilisateur. La plupart des changements auront lieu dans la partie inférieure de la structure hiérarchique qui contient les données relevant des domaines particuliers.

Le lexique conceptuel

On peut définir le lexique conceptuel comme l'organisation des concepts qui se trouvent à la base de la sémantique du langage naturel (d'après Pustejovsky 1990 : 2).

Les concepts sont le reflet du modèle du monde et sont soit de simples projections d'un type, soit le résultat d'une combinaison de différents types. Par exemple, le concept de « brique » est en fait la conceptualisation de la conjonction des types « artefact » et « construire ».

Le concept est représenté au moyen d'un graphe canonique, c'est-à-dire un réseau sémantique organisé autour d'une tête (Sowa 1984).

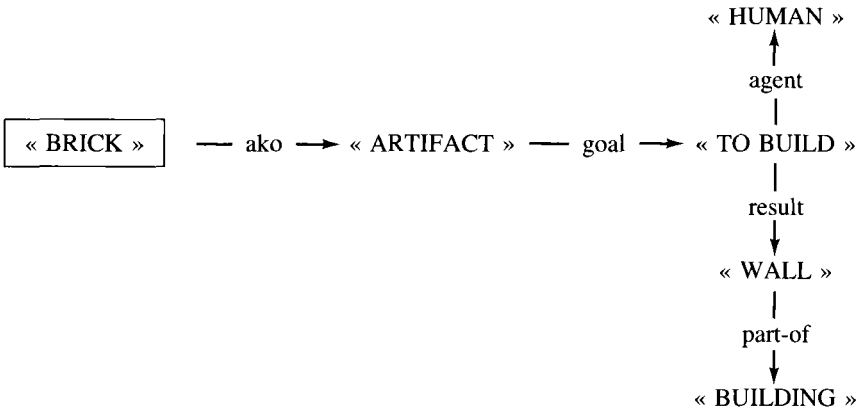


Figure 4

Ce réseau sémantique représente un concept, c'est-à-dire une unité de signification qui peut être définie au moyen d'une paraphrase, que nous appellerons *paraphrase définitoire*. Les arcs entre les nœuds ont une direction définie et sont étiquetés par des relations qui font partie de la définition du concept. Les relations indiquent comment deux nœuds sont reliés. Le concept sera réalisé par une unité lexicale lors du processus de lexicalisation (*cf. infra*).

Dans la traduction, le lexique conceptuel fonctionne comme une interlangue. Il est le pivot dans la recherche de la bonne traduction.

Le lexique conceptuel a donc pour but de rendre l'information sémantique explicite et accessible sous forme de réseaux sémantiques représentant les paraphrases qui définissent les concepts.

La paraphrase définitoire n'épuise pas la définition du concept. En effet, le concept est également associé à un certain savoir ajouté (*additional knowledge*) qui reste constant à l'intérieur d'une même culture. Les connotations et les métaphores reposent principalement sur ces connaissances ajoutées.

Le concept comporte deux parties :

a) la paraphrase définitoire, dont les éléments de sens sont hérités du modèle du monde comprenant les types et leurs relations;
b) le savoir ajouté.

Ces données sont reprises dans le schéma simplifié suivant :

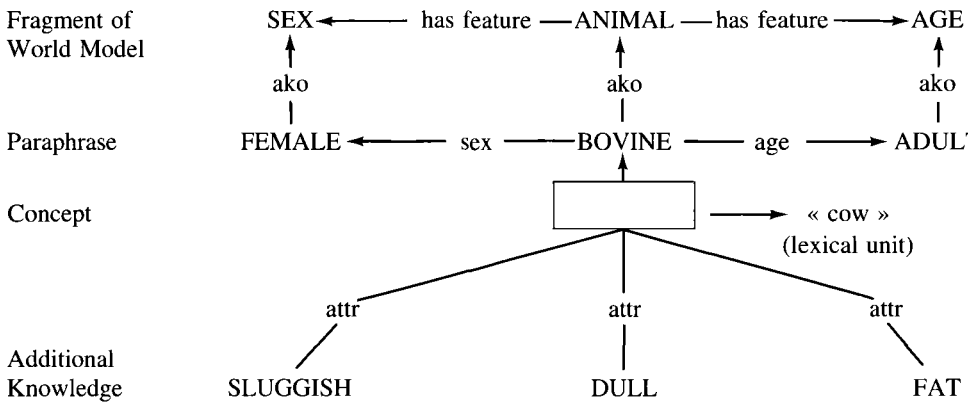


Figure 5

Notons que les concepts qui héritent d'un grand nombre de propriétés communes sont placés à un même niveau dans l'arbre hiérarchique, on peut en quelque sorte saisir le concept dans son *champ sémantique*.

Le lexique spécifique

Le lexique conceptuel met un concept en relation avec sa signification; il diffère du lexique particulier à chaque langue, qui lui met les concepts en relation avec les unités du vocabulaire. Le lexique conceptuel est donc une interlangue neutre par rapport à toute langue naturelle, mais il se trouve à la base du lexique spécifique de chaque langue. Le passage du lexique conceptuel au lexique spécifique est réalisé grâce au processus de lexicalisation qui fait correspondre chaque concept à un lexème défini.

Si un concept n'est pas lexicalisé dans une langue donnée, on peut néanmoins l'exprimer à l'aide d'une paraphrase. Cette paraphrase exprime en fait le premier concept lexicalisé supérieur dans la hiérarchie, auquel on ajoute les traits typiques manquants du concept non lexicalisé.

Par exemple, le lexème *embouteiller* est la lexicalisation de la construction conceptuelle formée des primitifs « mettre dans » et « bouteilles ». Ce lexème n'existe pas en allemand, il n'y a pas un verbe tel que « einflaschen ». On prend alors le concept lexicalisé supérieur et on y ajoute les traits typiques manquants de *embouteiller*.

En allemand, nous obtenons alors *in Flaschen* (trait typique) *einfüllen* (concept complexe supérieur). Le graphe conceptuel de *embouteiller* ressemblerait au schéma (simplifié) représenté ci-dessous.

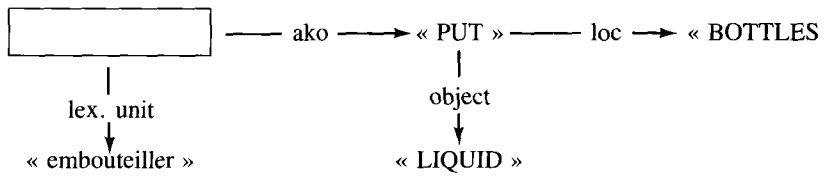


Figure 6

Les choses sont un peu plus compliquées lorsqu'on traduit dans le sens inverse, c'est-à-dire d'une langue où le concept n'est pas lexicalisé (*in Flaschen einfüllen* en allemand) vers une langue où le concept est lexicalisé mais où la paraphrase existe également (*embouteiller* ou *mettre en bouteilles/verser des bouteilles* en français). Lorsqu'il s'agit de traduction de textes écrits, nous prenons l'option de toujours choisir le terme le plus spécifique[3].

Un autre problème risque de se poser lorsqu'on veut modifier la paraphrase. Prenons le cas du lexème néerlandais *schimmel* paraphrasé en français par *étalon blanc*. Que se passe-t-il lorsqu'on veut traduire en néerlandais *un étalon blanc*? Dans ce cas, la modification porte en fait sur une partie seulement de la paraphrase, on dit de la couleur du cheval qu'elle est très blanche. Pour rendre cette notion de modification partielle en néerlandais, il faut faire éclater le concept et ne modifier que la partie concernée du concept, dans notre cas la couleur, ce qui donne *een heel witte schimmel*. Il ne s'agit pas ici d'un pléonasme puisqu'un *schimmel* n'est pas toujours **très** blanc. La modulation du modificateur et la décomposition du concept en sa paraphrase nous offrent donc une solution élégante à ce problème.

Conclusion

Nous avons émis l'hypothèse que le lexique conceptuel était un niveau clé pour la traduction automatique. Nous avons proposé de représenter ces connaissances sous forme de réseaux sémantiques, qui doivent permettre toutes les applications informatiques et le développement de logiciels d'aide à la traduction. À l'intérieur de ce formalisme, nous avons proposé des mécanismes pour choisir une traduction adéquate.

3. Cette même approche est défendue par Nogier (1991).

Bibliographie

AIT-KACI H. (1986) « An Algebraic Semantis Approach to the Effective Resolution of Type Equation », *Theoretical Computer Science 45*, p. 193-351.

BOGURAEV B. and J. PUSTEJOVSKY (1990) « Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design », *Proceedings of Coling '90*.

CARLSON L. and S. NIRENBURG (1990) *World Modeling for NLP*, Technical report CMU-CMT-90-121.

COSTERMANS J. (1980) *Psychologie du langage*, Bruxelles, Pierre Mardaga.

EMELE M. and ZAJAC R. (1990) *A Fixed Point Semantics for Type Systems*, Stuttgart, ms Polygloss.

EVENS M.W. (ed.) (1988) *Relational Models of the Lexicon, Representing Knowledge in Semantic Networks*, Cambridge University Press.

LEROT J. (1991) *Outline of the GENESE Model of Semantics Based Text Generation System*, Louvain-la-Neuve, ms Projet GENESE.

MEYER I., ONYSHKEVYCH B. and L. CARLSON (1990) *Lexicographic Principles and Design for Knowledge-Based Machine Translation*, Technical Report CMU-CMT-90-118.

NIRENBURG S. and V. RASKIN (1987) « The Subworld Concept Lexicon and the Lexicon Management System », *Computational Linguistics*, 13, 3-4, p. 276-289.

NOGIER J-F. (1991) *Génération automatique de langage et graphes conceptuels*, Paris, Hermès.

PUSTEJOVSKY (1989) *The Generative Lexicon*, ms. Brandeis University.

SOWA J.F. (1984) *Conceptual Structures*, Reading, Massachussets, Addison-Wesley Publishing Company.

SOWA J.F. (1988) « Using a Lexicon of Canonical Graph in a Semantic interpreter », in EVENS M.W. (ed.), (1988), p. 113-138.

THAYSE A. *et al.* (1988) *Approche logique de l'Intelligence Artificielle*, vol. 1, Paris, Dunod.

WIERZBICKA A. (1985) *Lexicography and Conceptual Analysis*, Ann Arbor, Karoma Publishers.

6

Pour une approche interactive au problème de la traduction automatique

Eric WEHRLI*

Département de linguistique générale et de linguistique française, Université de Genève, Genève, Suisse

Introduction

Au cours des quelques dernières années, on a pu observer un intérêt croissant pour des méthodes interactives dans le domaine de l'analyse du langage en général et dans celui de la traduction automatique en particulier. Contrairement aux systèmes non interactifs, ou autonomes, qui ne permettent aucune intervention de leur utilisateur en cours de processus, un système interactif est basé sur l'idée d'une collaboration entre un programme et son utilisateur. Autrement dit, ce type d'approche cherche à associer génie logiciel et génie humain plutôt que de les mettre en compétition. Concrètement, un système interactif tel que nous l'entendons ici est un système qui, lorsqu'il fait face à un problème qu'il est incapable de résoudre par lui-même, se réserve la possibilité de consulter son utilisateur pour, par exemple, lui demander de choisir entre différentes solutions ou lui permettre de corriger une éventuelle erreur. Les systèmes interactifs sont donc basés sur l'idée d'une collaboration directe, en temps réel, entre le progamme de traduction et son utilisateur, ce dernier conservant par conséquent un certain contrôle sur le déroulement du processus de traduction.

* Je tiens à remercier Robin Clark et Catherine Walther pour leurs commentaires et suggestions. La recherche décrite dans cet article est supportée en partie par un subside du Fonds national suisse de la recherche scientifique (subside no 11-25362.88).

Dans cet exposé, nous présenterons quelques-uns des arguments en faveur de systèmes interactifs dans le domaine de la traduction, et montrerons en quoi de tels systèmes s'opposent aux systèmes traditionnels de traduction automatique, non seulement du point de vue de l'utilisateur, mais également du point de vue plus technique qui est celui du concepteur. Nous aborderons ensuite certains problèmes liés au développement de systèmes interactifs et chercherons en particulier à répondre à des questions telles que :

- Quels types de problèmes devraient être traités de façon interactive?
- À quel moment est-il souhaitable d'engager le dialogue avec l'utilisateur?
- Comment convient-il de formuler les questions?

Pour une traduction interactive

La première question que je voudrais aborder ici est celle de la justification du recours à une approche interactive au problème de la traduction par ordinateur. En effet, d'aucuns pourraient objecter qu'il n'y a guère d'intérêt à développer des systèmes qui exigent de la part de leurs utilisateurs un travail non négligeable et probablement très astreignant s'il existe par ailleurs des systèmes de traduction capables de fonctionner de façon autonome. Autrement dit, pourquoi recourir à un mécanisme semi-automatique alors qu'il en existe d'entièrement automatiques?

Il est généralement admis que pour satisfaire les attentes de ses utilisateurs, la traduction par ordinateur doit parvenir à un haut niveau de qualité et de fiabilité. Or, sur la base des connaissances actuelles et de la technologie disponible dans les domaines de la linguistique théorique, de la linguistique informatique et de la traduction, ces exigences ne sont pas réalistes, et il n'y a guère d'espoir de les voir réalisées dans un futur proche par un système autonome[1].

Le processus de traduction repose sur un large éventail de connaissances, parmi lesquelles figurent en première place les connaissances grammaticales de la langue source et de la langue cible. Il est clair, cependant, que ces connaissances linguistiques sont insuffisantes pour assurer une traduction correcte et de bon niveau. Il est en effet souvent indispensable de disposer soit de connaissances plus générales, de nature encyclopédique, soit encore de connaissances spécifiques au domaine concerné par le texte. Cependant, alors que des progrès considérables ont été accomplis dans certains champs de la linguistique théorique, et en particulier dans les domaines de la syntaxe et de la morphologie, domaines pour lesquels il est maintenant possible d'envisager des applications informatiques relativement sophistiquées, d'autres aspects du langage et du processus de traduction restent fort mal compris. C'est en particulier le cas de la sémantique et de la pragmatique, pour lesquelles nous ne disposons que de modèles

1. Ces remarques ne s'appliquent pas aux applications qui touchent des domaines très limités et très spécifiques, comme le système TAUM-Météo (*cf.* Isabelle 1987) qui traduit de façon autonome et satisfaisante des bulletins météorologiques, système que Slocum (1985 : 8) décrit à juste titre comme « the world's only example of a truly fully-automatic MT system ».

très partiels. Quant aux modèles computationnels de représentation de la connaissance, ils n'existent pour l'instant que pour des domaines microscopiques[2].

L'incapacité des systèmes de traduction automatique actuels à assurer une qualité et une fiabilité suffisantes est largement reconnue, et dans la pratique, ces systèmes de traduction exigent un travail non négligeable de correction (et/ou de préparation) de la part de leurs utilisateurs. Dans des systèmes conçus comme des systèmes autonomes, cette intervention humaine, reconnue indispensable, s'exerce par nécessité sous la forme de la préédition et de la postédition. Dans le premier cas, l'utilisateur vient en aide au système de traduction en tentant d'anticiper les problèmes les plus visibles. Dans le deuxième cas, l'utilisateur essaie de corriger les problèmes les plus criants dans le document traduit, un travail souvent ingrat et peu apprécié des traducteurs professionnels. L'intérêt de l'approche interactive est de permettre à l'intervention humaine dans le processus de traduction de s'exercer non pas avant ou après la traduction d'un texte, mais pendant celle-ci, c'est-à-dire au moment où un problème surgit.

Quelques avantages des systèmes interactifs

Parmi les avantages que présentent les systèmes de traduction interactifs par rapport à des systèmes autonomes, le plus évident est sans doute le fait que l'utilisateur conserve un certain contrôle sur le déroulement du processus de traduction. Ainsi, face à des problèmes délicats d'ambiguïté, le système peut afficher les diverses possibilités et laisser le soin à l'utilisateur de sélectionner la plus appropriée d'entre elles. En effet, il arrive souvent que la sélection de la possibilité la plus satisfaisante repose sur des critères faisant intervenir des connaissances non linguistiques (connaissances encyclopédiques, bon sens, etc.). Une telle démarche présente le double avantage 1) de libérer le système informatique d'une tâche difficile, qu'il ne pourrait accomplir de façon satisfaisante, et 2) d'avoir recours au système expert le plus fiable et le plus performant en matière de compréhension du langage, l'homme.

Il arrive couramment que les textes à traduire contiennent des expressions peu claires ou ambiguës, même pour un traducteur humain connaissant bien le domaine. Il se peut également que l'information linguistique du texte source soit sous-spécifiée, en ce sens que certains traits linguistiques pertinents pour la langue cible ne figurent pas dans le texte source. Cela est particulièrement fréquent au niveau lexical, comme l'illustre le cas de l'anglais *river* qui correspond parfois à *rivière*, parfois à *fleuve*. Il est possible — mais non certain — que le texte contienne certains indices permettant de choisir le terme le plus adéquat, mais la recherche de ces indices et leur interprétation est une tâche pour laquelle l'homme est sans conteste mieux équipé qu'un programme. Un cas particulièrement extrême de sous-spécification est souvent cité dans la littérature. Il s'agit du problème du choix des déterminants dans la traduction du japonais vers l'anglais. En effet, comme les syntagmes nominaux en japonais ne portent pas de

2. Voir Carbonnel et Tomita (1987) pour une présentation d'un prototype de système de traduction basé sur la connaissance.

marque défini/indéfini, le traducteur doit déterminer pour chaque syntagme nominal s'il convient de l'exprimer en anglais sous la forme d'un syntagme défini ou indéfini.

Pour résoudre des problèmes de ce type, un système autonome devrait avoir accès à un énorme réservoir de connaissances, être capable de raisonnements très élaborés, et finalement compter sur la postédition. En l'absence de systèmes généraux de connaissances et d'inférences, les systèmes de traduction autonomes n'ont d'autre solution que d'utiliser des paquets de règles et de traits *ad hoc*, ce qui a pour conséquence d'augmenter la complexité des systèmes et d'en réduire la fiabilité. Un système interactif n'aura pas besoin de se soucier de ce type de problèmes puisqu'il peut se reposer sur l'intervention de l'utilisateur. En règle générale, chaque fois qu'un problème surgit qui dépasse les capacités du système informatique, celui-ci engage un dialogue avec l'utilisateur, à qui il incombe de résoudre le problème. Il s'ensuit qu'un système interactif tend à être plus réduit et plus simple.

Un autre avantage que l'on peut attendre d'un système de traduction interactif est la fiabilité, qui découle du fait que l'utilisateur conserve tout au long du processus de traduction une certaine maîtrise sur le processus de décision. En particulier, le fait que les problèmes les plus délicats sont directement soumis à l'appréciation de l'utilisateur limite considérablement le risque de propagation d'erreurs ainsi que celui de l'effet combinatoire des ambiguïtés. On appelle propagation d'erreur le fait qu'une erreur commise à un certain niveau dans le processus de traduction est fréquemment susceptible d'entraîner des erreurs à d'autres niveaux. La chaîne d'erreurs ainsi créée peut avoir des conséquences extrêmement dommageables et implique souvent une réanalyse complète de la phrase (ou du paragraphe) dans la phase de postédition.

Sur la base des avantages décrits dans cette section, on peut se demander pourquoi une telle approche n'a pas été développée auparavant. En effet, il peut apparaître bien surprenant que cette idée souvent discutée dans des « papiers académiques » et notamment dans Kay (1980), Tomita (1984, 1986), Johnson & Whitelock (1987) et Melby (1982, 1987), et plus récemment dans Zajac (1988), Boitet (1990), Jones & Tsujii (1990) parmi beaucoup d'autres, n'ait guère suscité de développements à grande échelle, à l'exception peut-être du système ITS décrit par Melby (1980, 1982). On notera, par contre, que de nombreux systèmes d'aide à la traduction ont été développés sur la base d'approches interactives, mais ces systèmes se limitent habituellement à des tâches simples, telles que la consultation de dictionnaires ou d'encyclopédies, ou encore la correction orthographique.

La raison, sans doute, est que le développement d'un système de traduction interactif, bien que plus petit et plus simple au niveau de son système de règles qu'un système autonome, reste une entreprise extrêmement complexe. De plus, comme nous allons le voir, les systèmes interactifs imposent des contraintes sévères à l'architecture du système de traduction ainsi qu'à la nature des algorithmes.

Structure et organisation d'un système interactif

Dans la section précédente, nous avons discuté un certain nombre d'avantages que les systèmes de traductions interactifs présentent sur leurs rivaux non interactifs. Au

nombre de ces avantages figurent 1) une plus grande fiabilité, 2) une relative simplicité (de leurs règles), puisqu'ils n'ont pas besoin de *tout savoir*, et 3) un garde-fou à la propagation et à la prolifération des erreurs. Nous allons voir, maintenant, que ces avantages ne sont pas gratuits, et que les systèmes interactifs imposent certaines contraintes et restrictions.

Tout d'abord, par définition, un système interactif nécessite un utilisateur pour l'assister, c'est-à-dire répondre à ses requêtes. Il s'agit là d'une contrainte majeure, probablement incompatible avec certaines applications de traduction. Cependant, cette contrainte peut sembler moins « drastique » lorsqu'on la compare aux tâches de prédédition ou de postédition des systèmes autonomes.

Du point de vue de l'utilisateur, une exigence fondamentale est que le niveau d'interaction soit raisonnable. En d'autres termes, les dialogues devraient être limités à un minimum. Les utilisateurs, en effet, abandonneraient sans doute rapidement un système qui poserait des questions à chaque mot!

La question de savoir quel taux d'interaction les utilisateurs considéreront comme acceptable est délicate et ne peut être résolue hors contexte. Il est clair que des paramètres tels que la complexité du texte à traduire, la qualité des résultats, le contexte même de la traduction entrent en compte, de même que des facteurs personnels.

Ce qui est sûr, cependant, c'est que les questions non pertinentes doivent être évitées à tout prix. Or, cela n'est malheureusement pas forcément facile. Pour comprendre pourquoi, commençons par considérer la question fondamentale du choix du moment le plus opportun pour engager une interaction. Les exemples donnés en (1) permettront d'illustrer le problème :

(1)a. John knew the answer.
b. John knew the answer was wrong.

Dans la phrase (1a) le syntagme nominal *the answer* fonctionne comme objet direct du verbe *knew*, tandis que dans l'exemple (1b) le même syntagme correspond au sujet de la phrase enchâssée *the answer was wrong*. Avant d'atteindre le verbe enchâssé, l'analyseur syntaxique, qui parcourt la phrase de gauche à droite et qui sait que le verbe *to know* peut prendre soit un objet direct, soit une phrase enchâssée, pourrait être tenté de lever cette ambiguïté immédiatement et, par conséquent, d'interroger l'utilisateur pour lui demander quelle option est la bonne dans ce contexte. Une telle question, bien sûr, est inutile, puisque l'ambiguïté disparaît d'elle-même lorsque l'analyseur arrive à la fin de la phrase, dans l'exemple (1a), ou lorsque l'analyseur lit le verbe *was*, dans l'exemple (1b).

On observera qu'en général il y a un conflit entre les exigences de l'utilisateur et celles du système par rapport à cette question du moment opportun de la consultation de l'utilisateur. Du point de vue de ce dernier, moins il y aura de questions, mieux cela vaudra, ce qui signifie que l'interaction devrait être différée le plus possible, de façon à pouvoir éliminer toutes les questions inutiles. Du point de vue du bon fonctionnement du système, par contre, il est souhaitable que l'interaction prenne place le plus rapidement possible, puisque c'est l'intervention de l'utilisateur qui va permettre de sélectionner les solutions les plus pertinentes et d'éliminer les moins pertinentes.

Une intervention rapide dans un cas d'ambiguïté permet d'éviter la propagation des erreurs et la combinaison des solutions.

À un autre niveau, il convient d'insister sur la nécessité d'un dialogue aussi convivial que possible. En particulier, il est important que les questions soient formulées de telle façon qu'un utilisateur non spécialiste de la théorie linguistique sous-jacente au système soit néanmoins capable de les comprendre. L'utilisation de fenêtres de dialogue et la présentation des diverses possibilités sous forme de menus dans lesquels l'utilisateur choisit une ou plusieurs réponses semblent répondre à cette attente.

Finalement, comme pour tous les systèmes qui travaillent en temps réel, il y a une contrainte d'efficacité. Il est en effet fondamental que les dialogues se suivent rapidement, de telle façon que l'utilisateur ne s'impatiente pas ou qu'il ne relâche pas son attention.

Il est intéressant de noter que ces contraintes ont des conséquences importantes sur l'architecture du système de traduction ainsi que sur le choix des algorithmes. Pour illustrer ce point, que nous ne pourrons développer en détail ici, bornons-nous à la question du moment de l'interaction dans la phase d'analyse d'une phrase. Il est clair que pour poser des questions aussi pertinentes que possible, l'analyseur doit non seulement noter la présence d'alternatives à un moment donné, mais également explorer dans une certaine mesure ces différentes possibilités, de façon à pouvoir garantir qu'elles ont une certaine pertinence dans le contexte en question. Comme, par ailleurs, il n'est pas souhaitable de différer toutes les questions jusqu'à ce que l'analyseur ait atteint la fin de la phrase, un certain degré de parallélisme dans le traitement des solutions semble nécessaire. De même, il apparaît que les deux stratégies classiques, ascendante et descendante, ne sont guère satisfaisantes et qu'une stratégie mixte semble préférable si l'on veut éviter les longues séquences d'hypothèses qui se révèlent finalement incompatibles avec les mots de la phrase à analyser (stratégie descendante) ou les nombreuses structures localement bien formées mais incompatibles avec leur contexte gauche (stratégie ascendante).

Niveaux d'interaction

Afin de rendre cette discussion un peu plus concrète, nous allons considérer dans cette section le cas d'un hypothétique système de traduction interactif et montrer les différents niveaux d'interaction. Nous admettrons que le système repose sur l'architecture classique avec les trois composantes d'analyse, de transfert et de génération. De plus, nous conviendrons que toute l'interaction est limitée aux deux premières composantes et, par conséquent, à une clarification du texte source.

Un premier niveau d'interaction prend place lorsque le texte source est considéré par l'analyseur, lors de l'analyse lexicographique. Le programme peut rencontrer des mots qui lui sont inconnus, c'est-à-dire des mots qui ne figurent pas dans ses dictionnaires. Il peut s'agir de noms propres ou de noms communs qui ne se trouvent pas dans ses dictionnaires, ou encore de mots mal orthographiés. Une fenêtre d'interaction est alors ouverte, exposant la phrase et mettant en évidence le ou les mots concernés.

Un menu permet à l'utilisateur de choisir l'action la plus appropriée parmi les possibilités suivantes :

- corriger la phrase;
- valider le mot comme nom propre;
- créer une nouvelle entrée de dictionnaire;
- interrompre le processus.

Un deuxième niveau d'interaction survient lors de l'analyse syntaxique, lorsque l'analyseur rencontre des ambiguïtés qu'il ne parvient pas à surmonter au moyen de ses règles et heuristiques. Cela concerne les cas typiques d'ambiguïtés syntaxiques, telles que les problèmes d'attachement illustrés en (2) :

(2)a. Jean frappe le chien avec le collier.
b. Jean frappe le chien avec les taches blanches.
c. Jean frappe le chien avec le poing.

Dans la phrase (2a), le syntagme prépositionnel *avec le collier* peut se rattacher soit au syntagme nominal *le chien* et alors il est interprété comme un modificateur de ce syntagme nominal, soit au verbe *frapper* et alors il est interprété comme un complément de ce verbe. La première interprétation est sans doute la plus naturelle dans le cas de la phrase (2b), la deuxième dans la phrase (2c). Dans des cas de ce type, la fenêtre de dialogue peut afficher la chaîne de caractères correspondant au syntagme prépositionnel à attacher et les têtes des syntagmes auquel il peut s'attacher. Dans notre exemple, cela reviendrait à afficher le syntagme prépositionnel [$_{PP}$ avec le collier] et demander s'il faut l'attacher au syntagme nominal objet direct *chien* ou au verbe *frappe*. Comme toujours, l'utilisateur a également la possibilité de modifier la phrase ou d'interrompre la session.

Un deuxième exemple d'ambiguïté syntaxique courante concerne l'interprétation de syntagmes antéposés, comme les syntagmes interrogatifs. Ainsi, dans la phrase (3) ci-dessous, *à qui* peut être interprété soit comme un complément du verbe *promettre* (*cf. Jean a promis à qui d'écrire*), soit comme un complément du verbe enchâssé *écrire*. La fenêtre d'interaction, dans un cas de ce genre, propose les possibilités, ainsi formulées :

(3) À qui avez-vous promis d'écrire?
a. écrire à qui?
b. promettre à qui?

Considérons un troisième cas classique d'ambiguïté syntaxique, cette fois en anglais.

(4)a. Flying planes can be dangerous.
b. Amusing people can be fun.
c. Visiting relatives can be boring.

Dans les phrases (4), les formes en *ing* (*flying, amusing, visiting*) peuvent être interprétées comme des adjectifs prénominaux modifiant respectivement les noms *planes, people, relatives*, ou comme des gérondifs, auquel cas les noms fonctionnent

commes des objets directs. La fenêtre de dialogue pourrait afficher les formes en *-ing* et les deux types d'interprétations, ce qui donnerait pour la phrase (4b) :

- *amusing* comme modificateur de [$_{NP}$ people];
- *amusing* comme verbe avec [$_{NP}$ people] comme complément.

Un troisième niveau d'interaction concerne les ambiguïtés d'ordre sémantique ou pragmatique, comme dans les exemples suivants :

(5)a. Jean ne boit pas parce qu'il est triste.
b. Jean dit à Paul que son livre se vend bien.
c. Les enfants se sont lavés.

La phrase (5a) illustre le problème des ambiguïtés de champ (ou de portée), c'est-à-dire du domaine sur lequel porte un quantificateur ou une négation. Ainsi, dans cette phrase, la négation peut porter soit sur le verbe seul, soit sur le verbe et la phrase circonstancielle. Dans le premier cas on obtient l'interprétation correspondant à la paraphrase *c'est parce qu'il est triste que Jean ne boit pas*, dans le second l'interprétation, très différente, correspondant à *ce n'est pas parce qu'il est triste que Jean boit.*

Dans de tels cas, il serait tentant d'afficher dans la fenêtre de dialogue les paraphrases correspondant aux diverses interprétations et de demander à l'utilisateur de choisir la plus appropriée. Cependant, le recours à des paraphrases peut être difficile et parfois conduire à de nouvelles ambiguïtés. Il est donc préférable dans de tels cas d'éviter les paraphrases et de formuler les questions comme suit :

- *parce qu'il est triste* est dans le domaine de la négation;
- *parce qu'il est triste* n'est pas dans le domaine de la négation.

Le deuxième exemple d'ambiguïté non syntaxique concerne l'interprétation des pronoms, comme dans l'exemple (5b). Dans cette phrase, le possessif *son* peut renvoyer soit à *Paul*, soit à *Jean*, soit encore à une autre personne mentionnée précédemment dans le discours. Ici, à nouveau, la fenêtre d'interaction va afficher la liste des antécédents potentiels et demander à l'utilisateur d'effectuer le bon choix.

Enfin, dans le cas de la phrase (5c), l'ambiguïté concerne l'interprétation du pronom anaphorique *se*. Dans cette phrase, *se* peut être compris avec un sens réfléchi ou réciproque. La fenêtre d'interaction peut afficher l'information suivante :

(6) Les enfants se sont lavés.
a. les uns les autres? (réciproque)
b. eux-mêmes? (réfléchi)

De nombreuses ambiguïtés se présentent au niveau du transfert, et notamment du transfert lexical, lorsque le système doit décider comment un élément lexical de la langue source doit être traduit. On sait que la correspondance entre items lexicaux de deux langues n'est pratiquement jamais une correspondance univoque. Autrement dit, il y a pratiquement toujours plusieurs possibilités de traduction pour un mot donné, même lorsque l'environnement syntaxique de ce mot a été bien dégagé. Ceci est en partie une conséquence directe de l'ambiguïté lexicale, dans la mesure où les différents sens associés à un mot sont susceptibles de traductions différentes. Il est légitime de penser que le dictionnaire bilingue contient toutes les correspondances possibles pour un item donné. On peut donc afficher cette liste dans la fenêtre d'interaction. Cette

façon de faire nécessite, cependant, une connaissance de la langue cible. Si, comme nous le souhaitons, l'interaction peut être limitée à la langue source, la sélection d'une correspondance se fera sur la base des différents sens du mot dans la langue source. On affichera alors le mot avec ses différents sens, en utilisant des synonymes. Par exemple, dans le cas (très simplifié) du mot *assurance*, on affichera:

- *assurance* syn. confiance;
- *assurance* syn. contrat.

Considérons un deuxième exemple, repris de Boitet (1990). Il s'agit de la traduction du mot *mur* dans une langue (comme l'espagnol) qui utilise des mots distincts pour le mur intérieur et le mur extérieur. La fenêtre de dialogue pourrait alors prendre l'allure suivante :

- *mur* — intérieur;
- *mur* — extérieur;
- ...

La sélection du sens le plus approprié d'un mot, qui permet au système de choisir une traduction correcte, est basée sur des exemples ou des synonymes associés aux entrées lexicales.

Conclusion

Les approches interactives au problème de la traduction par ordinateur présentent une solution de rechange particulièrement séduisante. En associant l'utilisateur au processus de traduction, les systèmes interactifs peuvent tirer parti des avantages respectifs de l'homme et de la machine. En outre, le fait que l'utilisateur conserve un certain contrôle sur le fonctionnement augmente la fiabilité et la robustesse du système. Enfin, alors que les systèmes autonomes nécessitent un long travail de postédition, tâche particulièrement délicate qui nécessite les compétences d'un traducteur professionnel, les systèmes interactifs du type discuté ici limitent l'interaction entre l'utilisateur et le programme à des dialogues en temps réel, habituellement sous la forme de menus dans lesquels l'utilisateur choisit un ou plusieurs items. De plus, si, comme nous l'admettons, des phrases complètement désambiguïsées peuvent être correctement traduites de façon automatique, il devrait être possible, en principe, de restreindre les interactions aux composantes d'analyse et de transfert et, par conséquent, de restreindre le dialogue à la seule langue source. La perspective de systèmes de traduction qui ne requièrent pas une connaissance de la langue cible n'est sans doute pas le moindre des avantages d'une telle approche.

Bibliographie

BOITET C. (1990) « Towards Personal MT », *Coling-90*, vol 3, p. 30-35.

CARBONELL J. and M. TOMITA (1987) « The Universal Parser Architecture for Knowledge-Based Machine Translation », *Proceedings of the 10th International Joint Conference on Artificial Intelligence*.

ISABELLE P. (1987) « Machine Translation at the TAUM Group », M. King (éd.), *Machine Translation Today*, Edinburgh University Press.

JONES D. and J. TSUJII (1990) « Interactive High-Quality Machine Translation for Monolinguals », *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Linguistic Research Center, Austin University of Texas.

JOHNSON R. and P. WHITELOCK (1987) « Machine Translation as Expert Task », S. Nirenburg (éd.), *Machine Translation : Theoretical and Methodological Issues*, Cambridge University Press.

KAY M. (1980) *The Proper Place of Men and Machines in Language Translation*, Technical report CSL-80-11, Xerox PARC.

KAY M. (1982) « Machine Translation », *American Journal of Computational Linguistics*, 8 : 74-78.

MELBY A., M. SMITH and J. PETERSON (1980) « ITS : Interactive Translation System », *Proceedings of the 8th International Conference on Computational Linguistics (COLING-80)*.

MELBY A. (1982) « Multi-level translation aids in a distributed system », *Proceedings of the 9th International Conference on Computational Linguistics (COLING-82)*, North-Holland.

MELBY A. (1987) « On Human-Machine Interaction in Translation », S. Nirenburg (éd.), *Machine Translation : Theoretical and Methodological Issues*, Cambridge University Press.

SLOCUM J. (1985) « A Survey of Machine Translation : its History, Current Status, and Future Prospects », *Computational Linguistics*, 11 : 1-17.

SLOCUM J. (éd.) (1988) *Machine Translation Systems*, Cambridge University Press.

TOMITA M. (1984) « Disambiguating Grammatically Ambiguous Sentences by Asking », *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84)*.

TOMITA M. (1986) *Efficient Parsing for Natural Language*, Kluwer Academic Publishers.

ZAJAC, R. (1988) « Interactive Translation : a New Approach », *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*.

7

La traduction automatique des bulletins d'avalanches de la Suisse

Pierrette BOUILLON, Katharina BOESEFELDT

ISSCO, Université de Genève, Genève, Suisse

Introduction

L'objectif de cette présentation est de montrer comment le système ELU (*Environnement Linguistique d'Unification*), développé à l'ISSCO, est utilisé dans le cadre concret de la construction d'un système de traduction automatique portant sur un domaine limité. Dans l'exposé, nous mettrons particulièrement l'accent sur la notion de représentation formelle et nous montrerons comment une riche représentation sémantique en structures d'attributs permet un transfert plus efficace entre les langues, ainsi qu'une approche moderne de la traduction automatique.

Après une brève présentation du projet, nous aborderons la mise en œuvre informatique avec le logiciel ELU, puis nous donnerons quelques exemples de représentation formelle en structures d'attributs.

Description du projet de traduction automatique des bulletins d'avalanches

Le projet

Le projet a pour but l'implémentation d'un système de traduction automatique capable de traduire des bulletins d'avalanches de la Suisse émis par *l'Institut Fédéral Suisse*

pour l'Étude de la Neige et des Avalanches (IFENA)[1] une à plusieurs fois par semaine pendant l'hiver. Rédigés initialement en allemand, ces bulletins doivent être traduits dans les deux autres langues officielles : le français et l'italien. Leur traduction automatique devrait permettre une traduction plus rapide, plus cohérente et plus systématique.

Le monde des avalanches : un sous-langage

Les bulletins d'avalanches constituent une application idéale pour la traduction automatique. Ils traitent du domaine limité des avalanches en Suisse. Ce domaine est bien connu et bien formalisé : des spécialistes ont défini les lois générales météorologiques, la manière de les exprimer et les zones géographiques de la Suisse (Salm 1986). Il est donc possible de définir le sous-langage utilisé pour la rédaction des bulletins d'avalanches et de limiter le traitement automatique à ce sous-langage. Or, dans un sous-langage, les ambiguïtés lexicales, syntaxiques et sémantiques sont réduites et plus aisément traitées. D'une part, le nombre de mots ainsi que leur sens sont restreints et les mêmes expressions reviennent à plusieurs reprises. D'autre part, les structures syntaxiques sont en nombre limité puisqu'il n'est pas nécessaire et même peu souhaitable pour des raisons de compréhension d'exprimer la même idée de différentes manières (Lehrberger 1982).

Unité sémantique des bulletins d'avalanches

Le but des bulletins d'avalanches est de donner une brève description d'ensemble des conditions de neige et de décrire le degré régional de danger actuel et prévu dans les Alpes suisses. Les bulletins sont généralement articulés en trois alinéas : la situation générale, les degrés du danger et l'indication des mesures à prendre et des dangers particuliers (annexe 1).

Chacune de ces trois parties utilise un vocabulaire bien défini et les bulletins sont donc rédigés au moyen d'un nombre limité de mots (environ mille), qui revêtent un nombre fini de significations et sont utilisés dans un nombre restreint de constructions. Pour vérifier la cohérence des syntagmes nominaux, il est assez aisé de définir un ensemble de traits sémantiques, de coder les noms et adjectifs avec ces traits et de stipuler que les mêmes traits doivent être récurrents à travers le syntagme. Le groupe nominal *le grand danger général*, par exemple, est correct, car le nom *danger* et les adjectifs *grand* et *général* ont reçu le même type sémantique **danger**.

De même, il est possible d'établir les unités de sens qui constituent les concepts de base du monde des avalanches et de les traiter comme des entités globales. Pour traiter, par exemple, le syntagme nominal français *couverture de neige* qui constitue un seul concept et qui, sémantiquement, doit être considéré comme une unité de sens,

1. Nous remercions l'IFENA qui subventionne partiellement le projet.

nous stipulons dans le lexique que le nom *couverture* requiert un argument qui a pour valeur *neige*.

Unité syntaxique des bulletins d'avalanches

La syntaxe, bien qu'elle soit, elle aussi, limitée, présente toutefois un degré de complexité assez intéressant, qui nous oblige à aborder des problèmes linguistiques parfois peu étudiés.

Contrairement aux notices météorologiques de TAUM-METEO (Lehrberger 1982), les bulletins sont en effet écrits à la troisième personne en prose et incluent des phrases simples ou subordonnées actives, passives, positives et négatives au présent, passé et futur, parfois coordonnées. Les groupes nominaux, adjectivaux, prépositionnels et verbaux sont souvent très complexes, avec plusieurs niveaux d'enchâssement, et ils sont souvent coordonnés.

Mise en œuvre du projet avec ELU : un environnement linguistique d'unification pour construire des systèmes de traduction automatique

Après avoir présenté le logiciel ELU, nous verrons la méthode de traduction qu'il utilise, puis nous mettrons en évidence l'originalité de ce logiciel. Enfin, nous aborderons l'état d'avancement du projet.

Le logiciel ELU

ELU (Estival 1989 et 1990, Russell 1990) est un programme qui a été développé à l'ISSCO et qui est capable d'interpréter des grammaires écrites dans un langage particulier. Inspiré de PATR II (Shieber 1986), il utilise le principe de l'unification pour combiner les informations, des ensembles de traits et de valeurs appelés structures d'attributs (voir exemple 1). Il possède trois modules principaux : **l'analyseur, le générateur et le module de transfert**.

L'analyseur permet d'obtenir pour une phrase les représentations en structures d'attributs permises par la grammaire.

Le générateur part d'une représentation de ce type et applique les règles de la grammaire pour découvrir la ou les chaînes de caractères qui peuvent être associées à la représentation en question.

Le module de transfert établit une relation entre deux mots ou deux représentations en structures de traits, pour permettre le passage d'une langue à une autre. Par exemple, la règle de transfert lexicale suivante met en correspondance directe les mots *Gefahr* et *danger* :

: TA : Gefahr danger

La règle de transfert structurale **pred** établit, elle, un lien entre les deux structures de traits < *head sem pred* > = *Valeur* de deux langues différentes L1 et L2 et lie la valeur X1 du chemin <* *head sem pred* > d'une langue L1 et la valeur X2 du même chemin dans l'autre langue L2 :

:T : pred
: L1 : <* head sem pred > = X1
: L2 : <* head sem pred > = X2
: X : X1 = X2

La traduction automatique avec ELU

La traduction avec ELU, fondée sur le modèle du transfert, se fait donc en trois étapes : la phrase en langue source est d'abord analysée pour produire une représentation en structures d'attributs; cette structure d'attributs est ensuite transformée par les règles de transfert dans une représentation conforme à la langue cible. Enfin, le générateur part de cette représentation pour générer le texte en langue cible.

Originalité de l'approche avec ELU

L'originalité de l'approche avec ELU réside dans la richesse de sa représentation en structures d'attributs, une génération et un transfert basés uniquement sur la représentation sémantique, la possibilité d'utiliser la même grammaire pour l'analyse et la génération, ainsi qu'un transfert bidirectionnel.

Représentation en structures d'attributs

Les ensembles de traits et de valeurs, appelés structures d'attributs (voir exemple 1), offrent un moyen flexible, élégant et efficace pour exprimer, de manière unifiée et condensée, toutes les informations jugées pertinentes par le linguiste, que ces dernières soient morphologiques, sémantiques ou syntaxiques.

Une génération et un transfert basés sur la représentation sémantique en structures d'attributs

Dans notre projet, la génération et le transfert se font uniquement sur la base d'une représentation sémantique, la syntaxe et les informations morphologiques non pertinentes pour la traduction et spécifiques aux langues n'étant pas prises en considération. Dans la sémantique, nous stipulons seulement quel est le prédicat de phrase, son ou ses arguments (dans une liste), le temps (présent, passé ou futur), si la phrase est positive ou négative et quels sont les modificateurs (voir exemple 1).

Utilisation de la même grammaire pour l'analyse et la génération

ELU permet d'utiliser la même grammaire pour l'analyse et la génération. Si l'utilisation d'une seule grammaire rend parfois plus complexe la tâche de celui qui écrit la grammaire, ce système offre cependant des avantages non négligeables. Tout d'abord, il ne nécessite pas l'écriture, puis le test fastidieux de deux grammaires différentes. De plus, il offre un moyen idéal pour vérifier si la grammaire écrite n'accepte pas des phrases illicites. Enfin, il assure la cohérence du système : toutes les phrases analysées peuvent aussi être générées par le système.

Un transfert bidirectionnel

Un des postulats d'ELU, qui découle de sa conception, est la possibilité de rendre le transfert réversible. S'il est possible de passer de la structure A à la structure B, étant donné la règle de transfert X, le contraire devrait aussi pouvoir être vérifié. En suivant cet objectif, nous pouvons donc tester les conséquences de cette approche moderne qui nous offre la possibilité d'être cohérents.

État d'avancement du projet

Après une étude fine du corpus avec l'aide de l'IFENA, l'accent est mis sur la traduction. Deux lexiques et deux grammaires du français et de l'allemand sont en cours d'élaboration et traitent de tous les phénomèmes de base du corpus des avalanches. Elles permettent l'analyse et la génération d'un grand nombre de phrases du corpus de complexité variable (annexe II). En même temps, nous rédigeons les règles de transfert qui permettent déjà la traduction bidirectionnelle de phrases du type de celles citées en annexe III. C'est de la représentation formelle des phrases qui figurent dans les annexes I et II qu'il sera question dans le chapitre cinq.

Exemples de représentation formelle des phrases en structures d'attributs

Comme nous l'avons dit plus haut, les phrases sont représentées par des structures d'attributs. Comme le monde des avalanches constitue un sous-langage où les variations structurales entre les langues sont peu nombreuses, nous avons essayé, lorsque c'était possible, de représenter les phrases allemandes et leur traduction française avec les mêmes structures sémantiques. Nous complexifions ainsi les grammaires d'analyse, mais nous simplifions considérablement les règles de transfert et nous gagnons ainsi en efficacité. Le transfert, en effet, est plus expérimental, encore moins bien connu et gagne à être limité.

Les exemples d'analyse suivants montrent la force d'expression des structures d'attributs, qui permettent d'exprimer le contenu sémantique des phrases. D'une part,

les structures de listes ([A,B]) et d'arbres (X(Y)) permettent une représentation très compacte. D'autre part, l'utilisation de structures rentrantes où deux traits possèdent la même valeur (#41 dans l'exemple 2) permet d'exprimer aisément des faits linguistiques complexes.

Exemple 1 (phrase active) :

in der Westschweiz besteht oberhalb 2000 m eine mässige lokale Schneebrettgefahr, traduit automatiquement par *un danger local modéré de plaques de neige subsiste au-dessus de 2000 m en Suisse Romande.*

```
args = [⟨81⟩ head : sem : compl : head : sem : compl : head : sem : compl = []
                                                                  detype = definite
                                                                  mod : n-comp : head : sem : pred = []
                                                                        nom = []
                                                                  pred = Schnee
                                                detype = non
                                                mod : n-comp : head : sem : pred = []
                                                      nom = []
                                                pred = brett
                                    detype = indefinite
                                    mod : alt : head : sem : pred = []
                                          expo : head : sem : pred = []
                                          n-comp : head : sem : pred = []
                                          nom : [⟨601⟩ head : sem : mod : degre : head : sem : pred = []
                                                                        pred = mässig
                                                ⟨602⟩ head : sem : mod : degre : head : sem : pred = []
                                                                        pred = lokal
                                          rel = []
                                    pred = gefahr ]
mod : alt : head : sem : mod : degre : head : sem : pred = []
                         pred = oberhalb (m(2000))
      loc : head : sem : compl = []
                         detype = definite
                         mod : expo : head : sem : pred = []
                               n-comp : head : sem : pred = West
                               nom = []
                               rel = []
                         pred = schweiz
      temps = []
morph : temps = present
        voix = actif
positif = yes
pred = bestehen
```

Cette structure d'attributs spécifie que la phrase est positive et au présent actif. La phrase possède comme prédicat le verbe *bestehen*, comme argument le groupe nominal sujet *eine mässige lokale Schneebrettgefahr* et comme modificateurs les groupes prépositionnels d'altitude et de lieu. Pour indiquer qu'il n'y a pas de groupe

prépositionnel de temps, nous utilisons une liste vide []. Pour que le transfert entre le français et l'allemand soit généralisable, nous avons décidé de décomposer les mots composés allemands dans le lexique pour obtenir la même représentation en allemand et en français. *Schneebrettgefahr*, par exemple, est décomposé en trois composants : une tête *Gefahr* avec un complément complexe composé d'une tête *Brett* et d'un complément *Schnee*.

Exemple 2 (phrase prédicative) :

in den Alpen ist die Gefahr noch gering und lokal, traduit automatiquement en français par *le danger est encore local et modéré dans les Alpes*.

```
args = [⟨238⟩ head : sem : args = [⟨38⟩ head : sem : ⟨41⟩
                                                      compl = []
                                                      detype = definite
                                                      mod : alt : head : sem : pred = []
                                                            expo : head : sem : pred = []
                                                            n-comp : head : sem : pred = []
                                                            nom = []
                                                            rel = []
                                                      pred = Gefahr]
                          mod : degre : head : sem : pred = []
                          pred = gering
        ⟨245⟩ head : sem : args = [⟨51⟩ head : sem  => #41]
                          mod : degre : head : sem : pred = []
                          pred = lokal]
mod : alt : head : sem : pred = []
      loc : head : sem : compl = []
                         detype = definite
                         mod : alt : head : sem : pred = []
                               expo : head : sem : pred = []
                               n-comp : head : sem : pred = []
                               nom = []
                               rel = []
                         pred = Alpen
      phrase : head : sem : pred = noch
      temps = []
morph : temps = present
        voix = actif
positif = yes
pred = und
```

Cette analyse montre un exemple d'utilisation de structures rentrantes pour exprimer le fait linguistique que les deux adjectifs coordonnés *gering und lokal* se rapportent au même nom *Gefahr* (#41). Dans les phrases prédicatives, le verbe *être* qui n'a pas de valeur sémantique n'apparaît pas dans la représentation. Le prédicat de la phrase qui est l'opérateur *und* possède deux arguments *gering* et *lokal* qui renvoient au même argument *Gefahr*.

Conclusion

Cet exposé montre une application concrète du logiciel d'ELU (*Environnement Linguistique d'Unification*). Après une brève description du projet, nous avons mis en évidence l'importance de la notion de sous-langage pour le développement d'un système de traduction automatique. La limitation du domaine nous a en effet permis la définition d'une représentation sémantique cohérente de la phrase pour un transfert efficace entre le français et l'allemand. Nous avons aussi montré que notre système de traduction automatique bénéficie de tous les avantages d'une approche nouvelle basée sur l'unification.

Bibliographie

ESTIVAL D. (1990a) *Elu User Manual, Technical Note 1*, ISSCO, Geneva.

ESTIVAL E., BALLIM A., RUSSELL G. and WARWICK S. (1990b) « A Syntax and Semantics for Feature-Structure Transfer », *The Third International Conference on Theorical and Methodological Issues in Machine Translation of Natural Language*, 11-13 juin 1990.

LEHRBERGER J. (1982) « Automatic Translation and the Concept of Sublanguage », *Sublanguage. Studies of Language in Restricted Semantic Domains*, edited by Richard Kittredge and John Lehrberger, Berlin, New York, de Gruyter, p. 81-106.

RUSSELL G., BALLIM A., ESTIVAL D. and WARWICK S. (1991) « A Language for the Statement of Binary Relation over Feature Structures », to appear in *Proceedings of European Association for Computational Linguistics*.

RUSSELL G., CARROLL J. and WARWICK S. (1990) « Asymmetry in Parsing and Generating with Unification Grammars : Case Studies from ELU », *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, June 1990, p. 205-211.

SALM B. (1982) *Lawinenkunde für den Praktiker*, Bern, Verlag des SAC.

SHIEBER S.M. (1986) « An Introduction to Unification Based Grammar », *CSLI Lecture Note*, n° 4.

Annexe I : Exemple de bulletin (9 mars 1984)

Starke Sonneneinstrahlung und eine Erwärmung zu Beginn der Woche sowie die nachfolgende kräftige Abkühlung haben die Lawinengefahr weiter gemindert. Die Niederschläge der letzten zwei Tage brachten lediglich dem Alpennordhang 10 bis 30 cm Schnee.

Am Alpennordhang, im Wallis in den Tessiner Bergen, in Nord- und Mittelbünden sowie im Unterengadin ist die Lawinengefahr dank einer stabilen Altschneedecke gering. Einzelne Gefahrenstellen können noch an Schattenhängen oberhalb rund 2000 m angetroffen werden.

Im Oberengadin und in den südlichen Bündner Tälern, wo das Schneedeckenfundament immer noch eine ungenügende Festigkeit aufweist, besteht für den Skifahrer vor allem an Nord- und Osthängen oberhalb rund 2000 m eine mässige örtlich beschränkte Lawinengefahr.

Annexe II : Exemples de phrases analysées et générées par la grammaire française

- Vingt à trente cm de neige fraîche sont tombés dans les Alpes, dans la région du Simplon, dans les vallées de la Viège et dans la région de la Maloja.
- Le rayonnement intense et l'élévation modérée de la température ont influencé le tassement de la couverture de neige.
- Les journées ensoleillées, mais froides ont influencé la stabilité de la couverture de neige.
- Les chutes de neige qui continuent causent une nouvelle aggravation du danger de avalanches.
- Une constitution défavorable de la couverture de neige incite à une prudence accrue sur les pentes raides.
- La couverture de neige présente des couches défavorables de givre qui incitent à la prudence.
- Le danger de plaques de neige est local et modéré dans les autres régions.
- Le touriste doit prendre garde à un danger élevé d'avalanches sur les pentes raides et ensoleillées où la couverture de neige est ramollie.
- Il faut considérer que la neige fraîche empêche le refroidissement de l'ancienne couverture humide.
- Bien que les chutes de neige soient insignifiantes, le danger local de plaques de neige augmente.
- Après que des avalanches sont descendues et qu'un refroidissement marqué s'est produit, la situation s'est calmée.
- Un danger local modéré de plaques de neige subsiste au-dessus d'environ 2000 m sur le versant nord des Alpes, dans la région du Gotthard, dans les Grisons et en Basse-Engadine.
- Le skieur doit éviter les pentes raides d'exposition nord à est.
- Les voies de communication sont menacées par des glissements de neige dans les bassins d'exposition sud.
- Les endroits dangereux se trouvent sur les pentes raides et sur les cuvettes d'exposition nord-ouest à nord-est.
- L'accès aux voies de communication et aux bâtiments est libéré.
- Après le temps variable des derniers jours, les hauteurs de neige atteignent environ 30 à 50 cm au-dessus de 1000 m sur le versant nord de les Alpes, en Engadine et en Valais après une semaine pauvre en précipitations.

Annexe III : Exemples de phrases transférées de l'allemand au français et du français à l'allemand

- *In den Alpen, in dem Simplongebiet, in den Vispertälern und in dem Malojagebiet fielen 20 bis 30 cm Neuschnee.*
- *Die Schneedecke betrug weniger als 10 cm.*
- *In den übrigen Gebieten beträgt der Neuschnee weniger als 10 cm.*
- *An dem Alpensüdhang fielen rund 80 cm Schnee.*

- *An dem ganzen Alpennordhang, in dem Wallis, in dem Gotthardgebiet und in dem Unterengadin besteht eine sehr grosse allgemeine Lawinengefahr.*
- *In dem Engadin besteht eine sehr grosse allgemeine Lawinengefahr.*
- *In den Alpen ist die Lawinengefahr noch gering und lokal.*
- *In den übrigen Gebieten ist die Schneebrettgefahr mässig und lokal.*
- *Der kräftige Temperaturanstieg und die nachfolgende Abkühlung beeinflussen die Schneedecke.*
- *An dem Alpennordhang und in dem Wallis betragen die Schneehöhen rund 20 bis 50 cm.*
- *Ein ungünstiger Schneedeckenaufbau ermahnt zu erhöhter Vorsicht.*
- *An dem Alpennordhang, in dem Wallis, in dem Gotthardgebiet und in dem Engadin besteht oberhalb 2000 m eine mässige lokale Schneebrettgefahr.*
- *In den übrigen Gebieten herrscht oberhalb 1800 m eine mässige lokale Schneebrettgefahr.*
- *Der wechselhafte Temperaturverlauf fu5hrte zu einer weiteren günstigen Setzung und Verfestigung der Schneedecke.*
- *Ein starker Temperaturanstieg und windbegleitete Schneefälle haben im Wallis zu einer Verscha5rfung der Lawinengefahr gefu5hrt.*
- *Die starke Erwärmung und die leichtere Abkühlung führten zu einer günstigen Setzung und Verfestigung der Schneedecke.*

8

Le projet EUROTRA, projet européen de recherche sur la traduction automatique

Christiane JADELOT, Dominique BOUSSEL

CELTA, CNRS / EUROTRA F-Nancy, Université Nancy 2, Nancy, France

EUROTRA

Finalité/originalité

Pour faire face à l'immense difficulté de traduction que représente le multilinguisme de la CEE (en même temps sa richesse), la Commission a lancé un projet de traduction automatique.

La Communauté européenne compte en effet 9 langues officielles : le danois, l'espagnol, l'allemand, le grec, l'anglais, l'italien, le néerlandais, le portugais et le français, ce qui fait 72 paires de langues si l'on veut les traduire toutes l'une vers l'autre.

Actuellement un certain nombre de documents ne sont pas traduits étant donné le coût très élevé du multilinguisme.

Face à cette situation la Commission a donc décidé de développer des outils d'aide à la traduction, le projet EUROTRA en fait partie.

Un prototype

Il s'agit en fait d'une maquette capable de traiter toutes les langues communautaires, et considérée comme une action de stimulation de la recherche visant à créer une

masse de connaissances en matière de traduction automatique (TA) et de linguistique informatique.

Au terme du programme on pourrait disposer d'un système opérationnel dans un domaine limité pouvant être développé ultérieurement dans la phase industrielle du projet.

Le système devrait couvrir des dictionnaires de 20 000 entrées pour chaque langue en 1992.

C'est en fait le premier programme communautaire de recherche (200 chercheurs) et de développement dans le domaine du traitement des langues naturelles contribuant :

- à la diffusion des connaissances et du savoir-faire,
- à la création d'infrastructures de recherche dans les pays membres,
- à la coopération européenne et à la prise de conscience d'une recherche communautaire dans ce domaine aussi importante que les objectifs techniques.

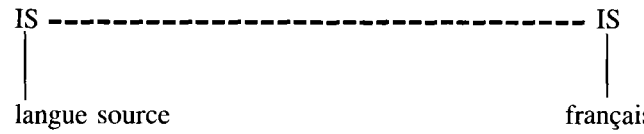
Le processus de traduction

Il se décompose en trois phases : l'analyse, le transfert et la synthèse.

L'analyse

Dans EUROTRA l'analyse, ainsi que la génération (synthèse) d'ailleurs, est monolingue. Elle est la même quelle que soit la langue cible. Cette caractéristique est motivée par le fait que le système traite un grand nombre de langues.

Il y a 9 modules d'analyse et 72 modules de transfert.



L'analyse d'une phrase repose sur une approche stratificationnelle très classique. Elle s'effectue selon quatre étapes :

1. analyse morphologique **EMS** : chaque mot de la phrase est étiqueté selon sa catégorie de discours;
2. analyse syntaxique **ECS** : les catégories du discours sont organisées en constituants tels que groupe nominal GN, groupe verbal GV, groupe prépositionnel GP;
3. analyse relationnelle **ERS** : les constituants sont réorganisés selon leur fonction syntaxique : verbe, sujet, objet, modifieur;

4. analyse sémantique **IS** : les fonctions syntaxiques sont organisées selon leur rôle sémantique : prédicat (gov), argument (arg) et modificateur (mod).

Chaque niveau d'analyse produit un arbre décoré. Le passage d'un niveau à l'autre est assuré par un transducteur. Le transfert d'un niveau à l'autre consiste en une modification d'arbres (ou élimination). Du point de vue du « framework » il n'y a pas de différence entre un transducteur du module monolingue et le transfert d'IS à IS. Chaque niveau inclut un dictionnaire et une grammaire.

La génération

Elle s'opère en effectuant les étapes dans l'ordre inverse.

IS → ERS → ECS → EMS → phrase cible

Les grammaires et les dictionnaires sont communs à l'analyse et à la génération, les modules sont dits « réversibles ». Seuls les transferts diffèrent.

Le formalisme

En se fondant sur des mécanismes d'unification, EUROTRA tente de créer un nouveau formalisme, l'hypothèse de base étant que le processus de traduction peut être décrit comme une suite de projections, de telle sorte qu'un texte subit de l'entrée initiale à la sortie finale, un certain nombre de représentations grammaticales. Ces représentations sont déterminées par des critères linguistiques, qu'ils soient morphologiques, syntaxiques en surface ou en profondeur, et sémantiques.

La représentation IS

Elle est basée sur une grammaire de dépendance. La structure linguistique est analysée au moyen d'éléments noyaux (gov) et de dépendants, soient les arguments (arg) soient les modifieurs (mod).

Prenons une phrase s1 en allemand et sa traduction t1 en français, différentes dans la forme de surface :

s1 : Die Kommission verliert die Angst vor dem Rat.
t1 : La Commission cesse d'avoir peur du conseiller.

Prenons les structures IS de ces phrases s1 et t1 :

Le transfert s'opère à partir de cette représentation. En examinant les deux représentations, on constate qu'elles sont **isomorphes**, c'est-à-dire que les arbres ont la même géométrie et ne diffèrent que par les traits qui introduisent les valeurs lexicales. Quelques règles lexicales et un certain nombre de données abstraites portées sur les nœuds de l'arbre rendent possible un transfert simple alors que les phrases sont structurellement différentes et permettent de générer la phrase cible.

Le transfert

Le choix d'un modèle à transfert plutôt qu'un modèle à langage pivot s'explique par le fait qu'un grand nombre de langues sont traitées ici. Il implique pour chaque langue l'écriture d'un module de transfert et d'un module de génération, étant donné que la représentation du texte source contient des informations sur la structure et le vocabulaire de ce dernier, ce qui n'est pas le cas d'un modèle à langage pivot qui lui est censé représenter le sens profond du texte source et procéder par concepts. Les langues traitées ici étant apparentées, des structures de la langue source peuvent trouver des équivalents dans la langue cible. Le degré d'abstraction de la représentation IS et son caractère sémantique sont les garants d'une traduction de qualité.

Le logiciel

Le logiciel appelé **ETS** (Eurotra Translation System, version 2.0.1.) fonctionne sur un ensemble de plates-formes **UNIX**™, il est écrit en langages **Prolog** et **C** et est construit autour d'un nombre de progiciels commerciaux (compilateurs Prolog, système de gestion de base de données, analyseur grammatical SGML, etc.). Le système complet, incluant le matériel linguistique, occupe une mémoire disque de 100 Mo. La configuration type est une machine UNIX™ d'une vitesse d'au moins 3 Mips CPU et dotée d'une mémoire centrale d'au moins 8 Mo.

Il est écrit par un groupe qui appartient à la Commission et il est distribué aux équipes nationales qui effectuent le travail linguistique.

EUROTRA a choisi de créer un langage spécifique à ses besoins, écrit en Prolog. Quelques notions comme l'unification et le *backtracking* sont typiques de Prolog mais le formalisme d'EUROTRA est conçu pour permettre l'écriture de règles de grammaire de manière plus simple qu'en Prolog.

Les composantes majeures du système sont les suivantes :

- un compilateur de règles;
- un interpréteur de règles;
- une base de données;
- un *text handling*;
- une interface homme-machine.

Le compilateur de règles

Il traduit les règles de grammaire écrites dans le formalisme propre à EUROTRA sous forme de clauses Prolog. L'avantage principal résultant du compilateur de règles est la vérification de la syntaxe. Un certain nombre de règles (dites *lexicon rules*) peuvent être intégrées à une base de données lexicales et ne sont plus dans les fichiers générateurs.

Le compilateur de règles est un programme écrit en C. Il utilise un analyseur lexical généré par **lex** et un analyseur grammatical généré par **yacc** (**lex** et **yacc** sont des outils d'UNIX™).

L'interpréteur de règles

L'interpréteur de règles réalise le transfert et la consolidation d'un objet d'un niveau à un autre en appliquant les règles de grammaire compilées qui alimentent ces niveaux.

Le transfert est réalisé par le transducteur, la consolidation par le générateur.

Pour produire la représentation d'un objet à un niveau donné, à partir de sa représentation au niveau précédent, le transducteur prend chaque nœud de l'arbre de départ, en partant de la racine, lui applique les règles de transfert puis le générateur consolide et affine ce nœud, avant de passer au nœud suivant. Chaque nœud sera évidemment décomposé à son tour, s'il n'est pas une feuille de l'arbre et chacune de ses parties sera transférée et consolidée.

L'interpréteur de règles fait partie du cœur du logiciel Prolog.

La base de données

La base de données **EDB** (Eurotra Data Base) est utilisée pour stocker les règles lexicales qui composent les dictionnaires monolingues du système. Pour le moment, la composante base de données est toujours en cours de développement. Elle permettra dans le futur de stocker les règles lexicales des dictionnaires de transfert.

Une base de données est nécessaire pour un certain nombre de raisons :

- sans une base de données efficace, une plate-forme système raisonnable ne pourrait pas maintenir plus de 5000 entrées lexicales; le but d'EUROTRA est d'avoir un lexique d'environ 20 000 entrées par langue à la fin de 1992;
- l'administration du système est plus facile si les règles sont dans une base de données. Elle peut contenir, en plus des règles elles-mêmes, d'importantes informations annexes comme l'auteur des règles, la date de création, le numéro de version du système au moment de création de la règle, un commentaire, etc.;
- dans un environnement multi-utilisateurs, les fichiers sont peu pratiques, ils ne peuvent être mis à jour que par une seule personne à la fois; une base de données permet de mieux contrôler les accès multiples aux données;
- il est aussi plus efficace de ne compiler que les nouvelles règles entrées dans la base plutôt que le fichier entier.

L'implémentation actuelle est basée sur **UNIFY** ™, un système de base de données relationnelle, commercialisé par *Unify Corporation*, qui accepte le langage standard industriel de base de données **SQL**.

Le *Text Handling*

Le sous-système *Text Handling* est une interface entre le texte écrit et les objets que le système EUROTRA utilise de façon interne.

C'est un processus à deux étapes. À l'entrée du système au début de la phase d'analyse du texte, le processus est appelé segmentation. La première étape produit une forme canonique à partir du texte d'entrée avec les codes de traitement de texte remplacés par les codes de SGML. À cette étape les caractères nationaux (et autres caractères spéciaux) sont normalisés. La deuxième étape reconnaît les éléments tels que marques de ponctuation, nombres, codes, etc. et extrait les mots du texte générant un objet qui est transmis au système de traduction. À la sortie du système (à la fin de la phase de génération de texte dans le langage cible) le processus est appelé désegmentation, les deux mêmes étapes sont appliquées dans l'ordre inverse.

Le sous-système *Text Handling* est basé sur **SGML** (Standard Generalized Markup Language). SGML est une norme de l'**ISO** (International Standards Organisation).

L'interface homme-machine

C'est une interface entre l'utilisateur et le cœur du logiciel. Il y a actuellement deux interfaces dans le système, une interface pilotée par menus et une interface de commande. La première donne accès à toutes les fonctions principales du système. La seconde donne un contrôle supplémentaire à l'expert (le linguiste) qui veut utiliser les fonctions interactives de debogage.

L'interface homme-machine comprend un certain nombre de composantes :

- le système de menus,
- l'interpréteur de commandes,
- le système d'aide,
- les fichiers de paramètres.

Le système de menus

Le système de menus permet un accès facile au système. En utilisant les menus, un utilisateur promène ses objets linguistiques à travers les différents niveaux du modèle formel d'EUROTRA.

Les objets sont stockés dans des fichiers.

À chaque niveau, les objets résultats peuvent être vérifiés en imprimant les fichiers ou en utilisant le *pretty printer* qui met l'objet sous forme d'arborescence.

L'utilisateur peut envoyer des traitements en arrière-plan à partir du système de menus, cela lui permet de continuer à travailler (édition, visualisation, etc.) pendant que le travail en arrière-plan s'exécute. Le travail en arrière-plan est choisi parmi une liste de travaux qui ont été préalablement définis dans le fichier de paramètres de l'utilisateur. Le système de menus permet un contrôle absolu sur ces travaux, ils peuvent

être surveillés ou interrompus. Un fichier-journal est créé par le système et peut être visualisé à partir du système de menus.

Le système de menus est un programme écrit en C, qui utilise le module **curses**. **Curses** est une bibliothèque d'UNIX™ qui permet la gestion du fenêtrage de l'écran.

L'interpréteur de commandes

L'interpréteur de commandes est disponible à partir du système de menus. Il peut être également utilisé tout seul. Toutes les fonctions du système de menus sont disponibles à partir de l'interpréteur de commandes.

Un nombre de fonctions supplémentaires sont utilisables à partir de l'interpréteur de commandes :

- contrôle des objets en mémoire. Les objets peuvent être laissés en mémoire, au contraire du système de menus qui écrit les objets dans des fichiers;
- possibilité de suivre l'exécution de la traduction en regardant quelles règles sont appliquées et dans quel ordre; contrairement au système de menus qui permet de suivre l'exécution globale de la traduction, on peut demander à ne voir que l'exécution de certaines règles.

L'interpréteur de commandes est intégré dans le cœur du programme Prolog. Le programme Prolog analyse les commandes et les exécute directement.

Le système d'aide

Le système d'aide en ligne est disponible à partir du système de menus.

Il offre à l'utilisateur un ensemble de sujets qui sont couverts par le système. En choisissant un sujet, l'utilisateur peut lire la documentation sur ce thème.

Le système d'aide n'est pas disponible à partir de l'interpréteur de commandes.

Les fichiers de paramètres

Il y a deux fichiers de paramètres utilisés dans le système EUROTRA : le fichier de paramètres de l'utilisateur et le fichier de paramètres du système.

Le fichier de paramètres de l'utilisateur sert à mémoriser toutes les informations concernant l'environnement de travail. Par exemple, les fichiers et répertoires de travail, le nom de l'éditeur que l'utilisateur préfère utiliser, etc. L'utilisateur peut changer ces valeurs.

Le fichier de paramètres du système est mis à jour par l'administrateur système. Ce fichier contient des informations qui en général changent seulement quand une nouvelle version du logiciel est disponible. Il contient, entre autres, les chemins d'accès aux utilitaires servant au logiciel.

Les fichiers de paramètres sont des fichiers texte. Chaque ligne commence par une lettre qui définit le type du paramètre, suivi par le nom du paramètre lui-même, suivi par sa valeur.

Extrait d'un fichier de paramètre du système :

VE, ets201
P, PROLOG, /usr/eurotra/rel201/bin/et-prolog
P, PR-PAR, « -h 1500 -g 512 -l 512 -a 256 -t 64 »

La première ligne définit la version du logiciel. La seconde donne le chemin d'accès à l'interpréteur Prolog et la troisième donne les paramètres avec lesquels l'interpréteur Prolog doit être appelé.

Les perspectives

Nous sommes actuellement dans une phase de deux ans dite de transition du programme EUROTRA qui a pour but une réflexion sur l'état de l'art et sur les développements futurs du produit obtenu.

En matière de recherche communautaire, bon nombre de problèmes linguistiques n'ont pas été résolus, voire même abordés (problèmes épineux de la traduction du temps et de l'aspect, du traitement des composés des langues germaniques dans les dictionnaires, de l'anaphore...); il nous semble donc que ces axes de recherche peuvent être maintenus, et cela en coopération avec de futurs utilisateurs, lexicographes, traducteurs.

Puisque le thème de ces journées est *l'environnement du traducteur en l'an 2001*, nous aimerions revenir, en guise de conclusion, sur certains aspects techniques du logiciel et les possibilités qu'il peut offrir à un traducteur.

En effet, après la période d'euphorie des années soixante où l'on pensait que la machine traduisante pourrait tout faire, un certain nombre de recherches dans ce domaine ont été abandonnées.

Par la suite, des travaux menés sur des applications ont amené les chercheurs à voir les choses de manière plus modeste et à concevoir non plus la traduction comme automatique mais comme un outil d'aide au traducteur, efficace et souple. Nous pensons que l'avenir d'un programme tel qu'EUROTRA se trouve dans le développement de l'outil dont nous disposerons à la fin de 1992, en mettant l'accent sur sa souplesse et sa grande maniabilité y compris pour un novice; parmi les outils à développer, généralisables à d'autres applications de traduction automatique ou de traitement du langage naturel on citera :

- l'outil *text handling*,
- l'outil grammaire,
- l'outil lexique,
- l'outil traitement linguistique.

L'outil *Text handling*

Il permettra de manipuler des textes (sections, paragraphes, tableaux, notes de bas de page), le stockage de textes et outils documentaires permettant de produire des concordances, des fréquences, l'analyse orthographique et la segmentation de texte.

L'outil grammaire

Il consistera en plusieurs outils pour ceux qui écrivent les grammaires : codage de la grammaire, possibilité pour l'utilisateur d'écrire et de maintenir les grammaires dans le formalisme d'EUROTRA, et cela grâce à des éditeurs spécialisés et des menus; possibilité de définir des sous-grammaires pour des tests spéciaux ou des applications particulières.

L'outil lexique

Il comprendrait plusieurs outils pour des lexicographes et des linguistes : codage du lexique, possibilité d'éditer des entrées lexicales, le codage étant assisté par un menu, dans le même esprit que l'actuelle base de données; importation de lexiques; possibilité de définir des sous-lexiques; vérification de la cohérence du lexique.

L'outil traitement linguistique

Il consistera en plusieurs sous-modules qui aideront l'utilisateur durant les phases d'essais et de débogages :

- la machine virtuelle fournira les capacités d'un moteur d'inférences (unification, résolution de contraintes). Les opérations exécutées sont l'analyse grammaticale, l'évaluation des contraintes et la génération;
- un débogueur offrira plusieurs possibilités pour superviser les opérations de la machine virtuelle;
- la possibilité d'examiner des sous-résultats (texte ou graphique);
- un agent de transfert permettra à l'utilisateur de spécifier une application de traduction automatique basée sur le transfert.

La multiplicité des langues traitées par le système (72 paires) a imposé une stratégie de transfert simple à partir d'une représentation IS relativement abstraite en regard de la structure de départ. Ce passage obligé est la condition pour une traduction de qualité satisfaisante. De nombreuses et longues recherches sont encore nécessaires pour mettre au point cette représentation car les études menées de manière contrastive montrent que chaque fois qu'une information linguistique se réalise différemment d'une langue à l'autre, une représentation abstraite doit être mise au point, cohérente avec tous les phénomènes qui ont été calculés auparavant. Ceci demande un vaste programme de

recherche pour faire aboutir l'état des connaissances sur les neuf langues. À court terme on peut espérer que le système produira des traductions laissant à l'humain le choix de la traduction correcte, à long terme nous souhaitons que le programme EUROTRA aura montré que les recherches sur l'interface sémantique et sa portée interlinguistique doivent être poursuivies.

Bibliographie

BECH A. (1990) « The Virtual Machine », *Eurotra Reference Manuel 6.1*, Technical Report of Eurotra, Luxembourg.

DANLOS L., BECH A., CAROLI F., DAILLE B., MESLI N., NAMER F., NOHR S. (1989) « Support Verbs and Predicative Nouns in a Machine Translation System », *Eurotra Reference Manuel 6.0*, Technical Report of Eurotra, Luxembourg.

DANLOS L., LAURENS O. (1991) « Présentation du projet Eurotra et des grammaires d'Eurotra-France », *Rapport technique Eurotra-Talana*, n° 1, Paris.

DANLOS L., NAMER F. (1991) « Les constructions à verbe support : propriétés linguistiques et implémentation dans Eurotra », *Rapport technique Eurotra-Talana*, n° 4, Paris.

EUROTRA (1991) « Eurotra Reference Manual 7.0 », *Technical Report of Eurotra*, Luxembourg.

GAJEK O., LUKS K., MEYER R., NETTER K., SCHUETZ J., THURMAIR G., WEHRMEYER A. (1990) « Requirements Study for a EUROTRA-II Software Environment », *Internal Progress Report*, Luxembourg.

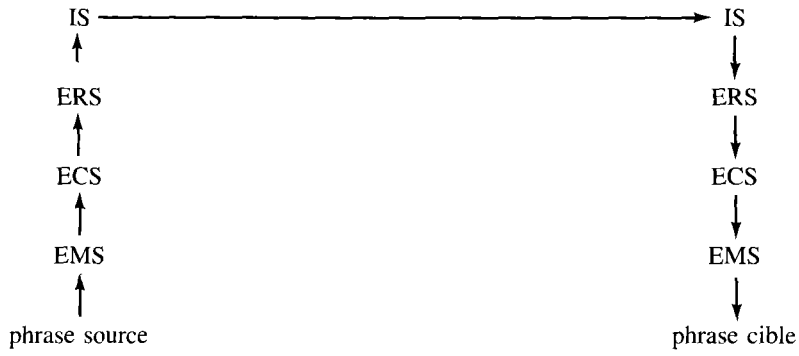
GAJEK O., LUKS K., MEYER R., SCHUETZ J., THURMAIR G., WEHRMEYER A. (1990) « Assessment Study for a EUROTRA-II Software Environment », *Internal Progress Report*, Luxembourg.

NAMER F. *et al.* (1990) *Rapport d'implementation Eurotra-France*, Paris.

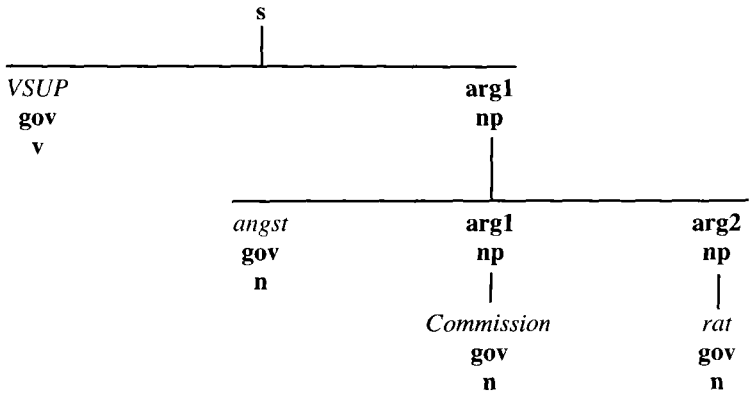
SCHUETZ J. (1990) « A Multi-Agent Architecture for a NLP Development Environment », Ms., IAI, Saarbrücken.

SCHUETZ J. (1990) « An Architecture for Reusable Lexical Resources in a Multi-theoretical Environment », Ms., IAI, Saarbrücken.

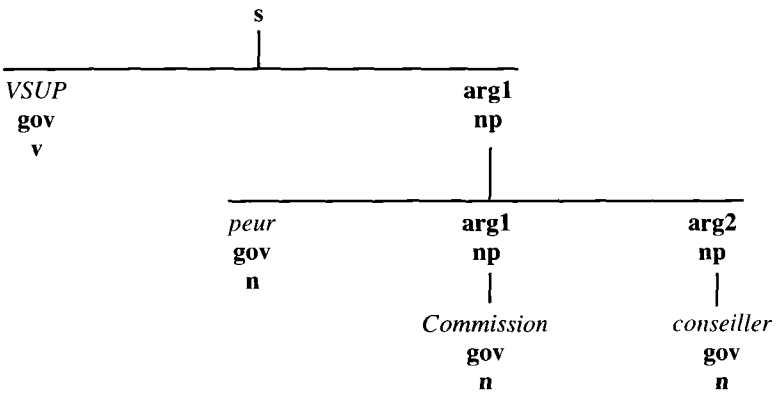
Annexes


IS
IS
ERS
ERS
ECS
ECS
EMS
EMS
phrase source
phrase cible

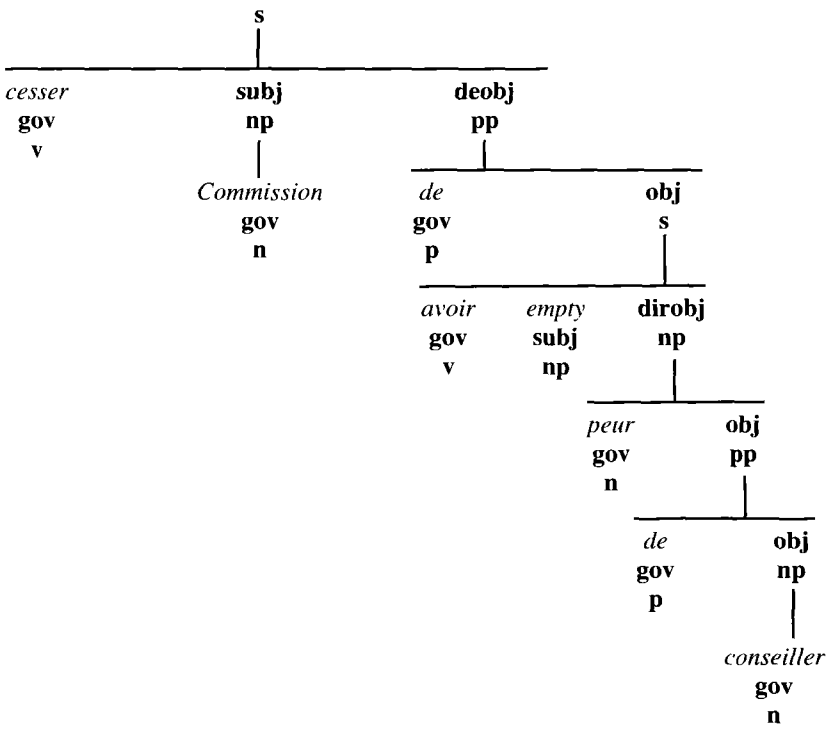
ISD


s
VSUP
gov
v
arg1
np
angst
gov
n
arg1
np
Commission
gov
n
arg2
np
rat
gov
n

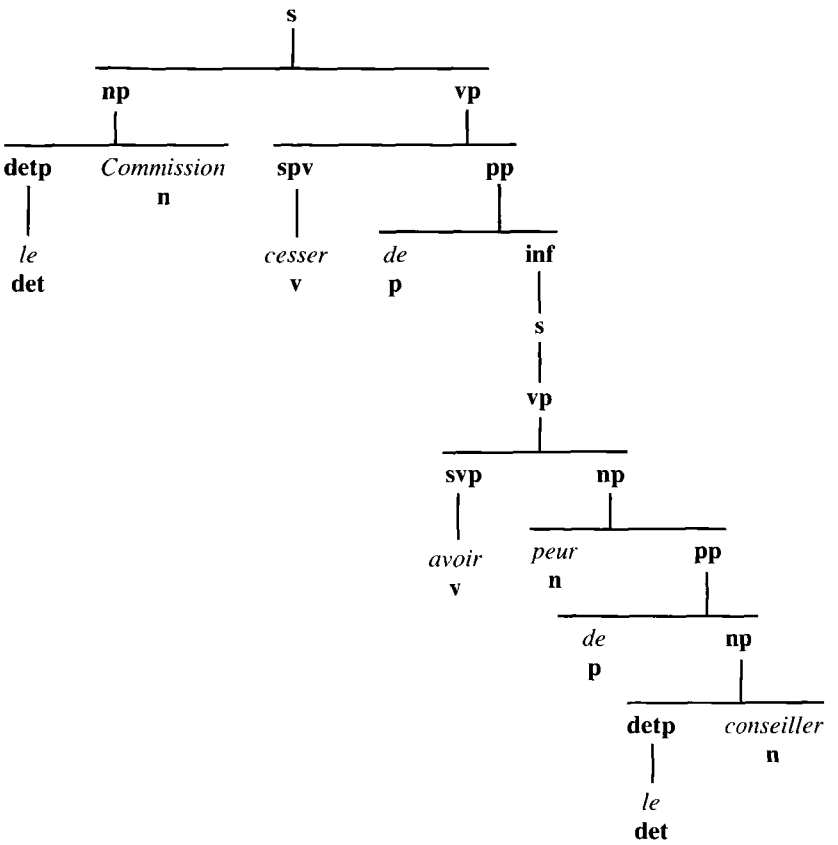
ISFR


s
VSUP
gov
v
arg1
np
peur
gov
n
arg1
np
Commission
gov
n
arg2
np
conseiller
gov
n

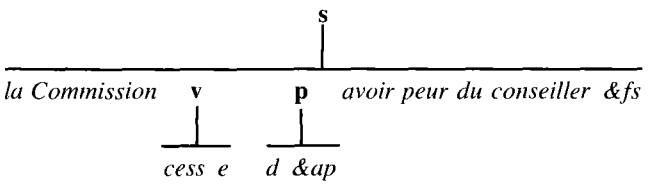
ERSFR


s
cesser
gov
v
subj
np
Commission
gov
n
deobj
pp
de
gov
p
obj
s
avoir
gov
v
empty
subj
np
dirobj
np
peur
gov
n
obj
pp
de
gov
p
obj
np
conseiller
gov
n

ECSFR



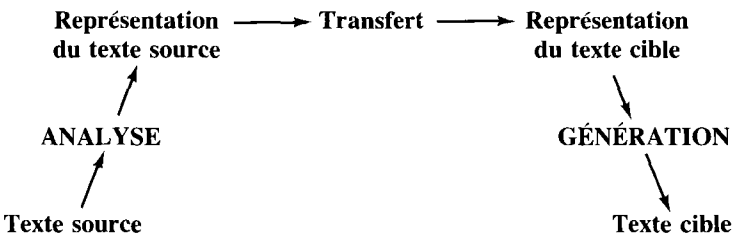
EMSFR

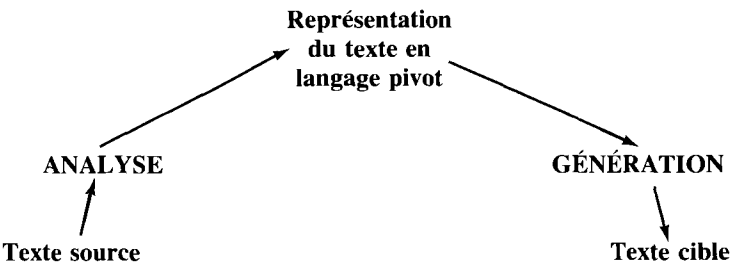


TEXTE

%% LANGID = fr
%% 1

la Commission cess e d avoir peur du conseiller.



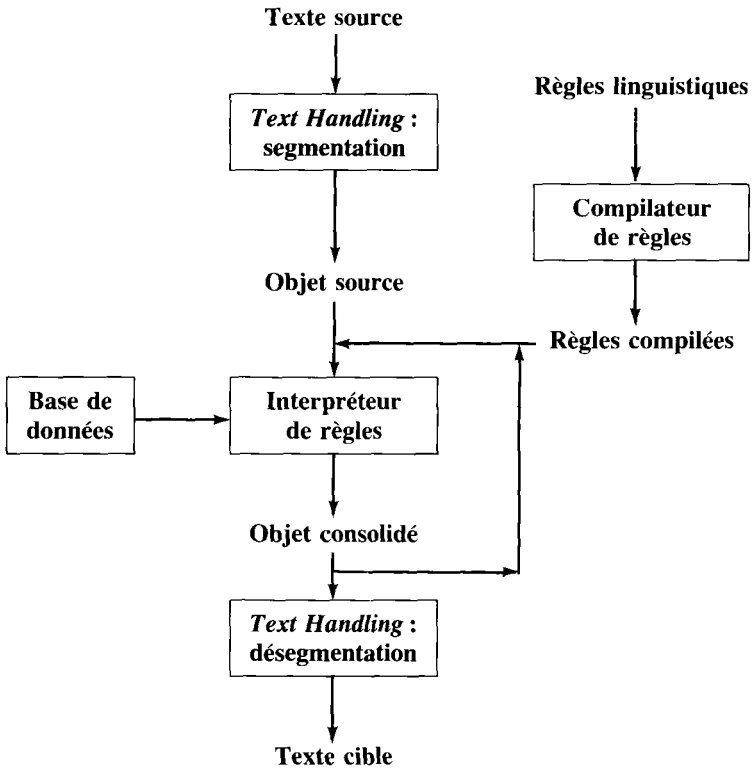


Exemple de phrase à traduire :

Heureusement, la Commission prend des mesures importantes.

Objet utilisable par l'interpréteur de règles, après segmentation :

```
:filetype:object.
language(text,fr).
obj(text,fb(cat,io,lex,rend),l:[]:{cat=s,io='0'}.[
l:[]:{lex='Heureusement',rend=no}.[]
,l:[]:{lex='&cm'}.[]
,l:[]:{lex='la',rend=no}.[]
,l:[]:{lex='Commission',rend=no}.[]
,l:[]:{lex='prend',rend=no}.[]
,l:[]:{lex='des',rend=no}.[]
,l:[]:{lex='mesures',rend=no}[]
,l:[]:{lex='importantes',rend=no}.[]
],label([unknownfile/(text/1),'phrase'])).
```



Eurotra Translation System ets v2.1.0.9 26 Feb 1991

Dataset : Batch Pid : Dir :

Ets Toolbox

0. Unify User Environment
1. Segment Input Text
2. Compile Rules
3. Translate Objects
4. Invoke Command Interpreter
5. Edit File
6. Review Paramaters
7. Object Handling
8. Batch Process Control
9. Document Processing

Enter your choice :

Help ^ Xit !Shell ^ Wash

9

ODILE 2 un outil pour traducteurs occasionnels sur Macintosh

Mathieu LAFOURCADE

GETA, Institut IMAG (UJF & CNRS), Grenoble, France

Résumé

La conception d'outils informatisés d'aide aux traducteurs occasionnels est délicate. En effet, il faut intégrer des outils déjà existants et interagir avec les applications du commerce. Nous avons réalisé une application sur Macintosh, Odile, qui permet, en utilisant un lemmatiseur, de fabriquer à partir d'une chaîne sélectionnée un ensemble de clés de recherche, de les compléter si nécessaire « à la main », et de les utiliser pour accéder à des dictionnaires usuels supportés par un utilitaire du commerce. L'architecture d'Odile est modulaire. Elle repose sur la définition d'interfaces génériques avec des lemmatiseurs et des outils dictionnaires. Après avoir présenté la version actuelle d'Odile, on indique comment l'améliorer. En particulier, l'introduction d'applications communicantes au niveau du système d'exploitation devrait permettre la réalisation de fenêtres « vocables », et l'élimination presque totale des interventions de l'utilisateur.

Introduction

Dans le cadre de la THAM (Traduction humaine assistée par la machine), nous nous intéressons à la création d'outils destinés aux traducteurs occasionnels. Il ne s'agit donc pas d'outils pour professionnels, qui peuvent être dédiés, mais de possibilités supplémentaires offertes aux utilisateurs de traitements de texte, tableurs, SGBD et autres outils du commerce. C'est pourquoi nous n'envisageons pas pour l'instant

d'accès à des « bases de bitextes » résultant de traductions antérieures (comme TSS d'ALPS sur IBM [8]).

Ce genre d'outil est destiné à un public non spécialisé (ni informaticien, ni linguiste), et vise à automatiser le processus de traduction. Si l'on estime qu'environ trente pour cent du temps de traduction est passé à chercher dans des dictionnaires, grammaires et autres, il semble clair qu'une informatisation de ce genre de tâches augmente la productivité. Les dictionnaires de tels outils doivent être, au moins partiellement, modifiables par l'utilisateur. Il faut aussi éviter de forcer l'utilisateur à taper une « forme canonique » (« cheval » pour « chevaux »). Pour l'instant, seuls des outils dédiés comme SISKEP [14] ont ces deux possibilités. Il est donc intéressant de les offrir à l'intérieur d'un outil ouvert. D'autre part, ce type d'outil peut servir à d'autres choses qu'à la traduction, les dictionnaires pouvant aussi bien être des fiches techniques ou des thesaurus.

La première partie présente de manière plus précise les motivations, les désirs et les contraintes que nous avons pris en compte pour la spécification d'Odile. La partie centrale de l'exposé est consacrée à la description de la version 2 d'Odile, une application sur Macintosh qui intègre un lemmatiseur et un outil dictionnaire. Odile permet, en utilisant un lemmatiseur, de fabriquer à partir d'une chaîne sélectionnée un ensemble de clés de recherche, de les compléter si nécessaire « à la main », et de les utiliser pour accéder à des dictionnaires usuels supportés par un utilitaire du commerce, WinTool™. Sa réalisation a été l'occasion de définir une interface universelle pour les lemmatiseurs. L'essentiel de la première version a été réalisé dans le cadre du mémoire CNAM d'Isabelle Tomasino [13]. La version 2 correspond à une nouvelle spécification de l'interface homme-machine, de l'interface logicielle et des fonctionnalités. La dernière partie montre un certain nombre d'améliorations possibles. En particulier, l'introduction d'applications communicantes au niveau du système d'exploitation devrait permettre la réalisation de fenêtres « vocables », et l'élimination presque totale des interventions de l'utilisateur.

Spécifications

Motivations

Les problèmes avec les outils existants

L'inconvénient majeur des outils existants est qu'ils sont souvent fermés (cas de WordFinder sur Macintosh). Le traducteur occasionnel ne peut donc en modifier les dictionnaires ni les règles. L'utilisateur reste donc dépendant du contenu des bases fournies avec de tels outils.

D'autres outils sont largement dépendants d'un environnement proposé en même temps (SISKEP de l'USM [14] et [15] et TAIM extension de Palantir de Melby [9]). L'utilisateur est alors obligé d'apprendre l'utilisation d'un nouvel éditeur.

Des outils ouverts existent (Mercury [10] , WinTool [16]) mais imposent d'entrer une clé de recherche correcte. L'utilisateur ne peut pas rentrer des clés approximatives.

On constate donc qu'aucun des outils proposés n'est à la fois ouvert et indépendant d'un environnement. De plus, ils obligent l'utilisateur à taper une « forme canonique » pour effectuer la recherche.

Les clés de recherche doivent être générées automatiquement ou à la main

Pour le français, il est souhaitable d'avoir un système automatique qui élabore les clés correspondant à un morceau de texte. De plus, l'utilisateur doit pouvoir gérer des clés « personnelles », car un dictionnaire automatique ne sera jamais ni complet ni parfait. Il est nécessaire de pouvoir introduire des termes composés, par exemple pour trouver « manette de sécurité » à partir de « manette de » sans se retrouver sur « mangeoire » après tous les « manette de... ». Il est également pratique de spécifier des équivalents de sigles, par exemple qu'OTAN correspond à NATO et de ne pas avoir à taper « Organisation du Traité de l'Atlantique Nord ».

Un générateur automatique de clés doit rester simple. Par conséquent, on ne le munira pas de fonctions complexes permettant, par exemple, de retrouver « mettre les points sur les i » à partir de « dans les copies de ses élèves, elle ne mettait pas toujours et partout les points sur les i ».

Un champ d'expérimentation avec une interface conviviale

La manipulation classique dans les outils d'aide à la traduction est de laisser l'utilisateur sélectionner une clé et de la chercher dans un dictionnaire. Nous pensons que ce type d'interaction peut être amélioré.

Désirs

Accessibilité universelle (programme résident)

L'utilisateur veut pouvoir utiliser les logiciels dont il a l'habitude en même temps que des outils d'aide à la traduction. On désire donc un outil indépendant de toute application particulière (voir Mercury [10]).

Généricité

Nous voulons pouvoir adapter n'importe quel lemmatiseur ou outil dictionnaire. PILAF [4] est adapté aux langues romanes. D'autres lemmatiseurs seraient adaptés à d'autres types de langues. On pourrait également inclure des lemmatiseurs « à dictionnaires restreints » (Palmer [12]), munis de méthodes heuristiques pour le traitement des mots inconnus.

En tant que chercheurs, nous voyons aussi là un moyen de rendre accessible les produits de laboratoire à un grand nombre d'utilisateurs. En effet, développer un environnement convivial et fiable, ou bien adapter les données d'un outil à un autre, représente un travail de développement trop lourd pour des équipes de recherche.

Convivialité

L'utilisateur veut lancer un processus de recherche et ne plus s'occuper de rien. L'interactivité doit donc être faible. Cela peut se résumer par la formule suivante : « give and leave ». D'autre part, l'utilisateur doit en regardant avoir l'intuition de ce qu'il peut faire (« look and feel »). Un certain nombre de principes d'ergonomie doivent être respectés (Apple [1] et [2] et Coutaz [5]).

Enfin, la rapidité d'exécution est primordiale : sans elle, l'utilisateur préférera ses dictionnaires papier.

Contraintes

Nous avons choisi de réaliser Odile sur Macintosh, en intégrant un lemmatiseur tiré du système PILAF [3] et un outil dictionnaire disponible (WinTool™).

Le Macintosh

Cette gamme de micro-ordinateurs offre une interface conviviale. L'interface générale d'Odile suit les spécifications préconisées par Apple.

Pilaf

Pour des raisons de rapidité d'implantation, nous avons adapté un lemmatiseur déjà existant sur le Macintosh (Courtin *et al.* [4]). Le moteur de PILAF existait en langage Pascal sur micro-ordinateur (compatible PC). Un premier portage sur le Macintosh a été fait dans le cadre du PRC-CHM en 1988-89 avec un travail d'adaptation sur le lemmatiseur. Un dictionnaire d'environ 25 000 lemmes (produisant environ 175 000 formes de sortie) provenant du travail de l'équipe TRILAN. Il s'agit d'une adaptation du dictionnaire DELAS du LADL (M. Gross).

WinTool™

WinTool™ est un outil dictionnaire déjà existant dans le commerce. Pour l'intégration, nous avons eu accès au code. D'autre part, WinTool™ est très rapide, peu cher et ne prend pas beaucoup de place mémoire. En effet, la taille du code est d'environ 100 Ko et la taille des fichiers représente environ 50 octets par articles.

Odile version 2

Odile version 2 est une réécriture complète de l'interface et une redéfinition des fonctionnalités de la première version d'Odile [13]. Des démonstrations sont possibles, soit du français vers l'arabe (avec système arabe, coopération avec l'IERA dans le cadre RFIL), soit du français vers le malais.

Interface et fonctionnalités

Exemple de session

Comment se présente Odile? Plutôt que de décrire les unes après les autres les fonctionnalités de cet outil, nous présentons le déroulement d'un scénario type. Supposons qu'un chercheur doive traduire son dernier article en vue d'une présentation à un congrès. Pour cela, il utilise son traitement de texte favori et traduit les paragraphes les uns après les autres directement sur le document. Cependant, il hésite pour traduire le fragment « mettent les pieds dans le plat » (« (they) spill the beans »).

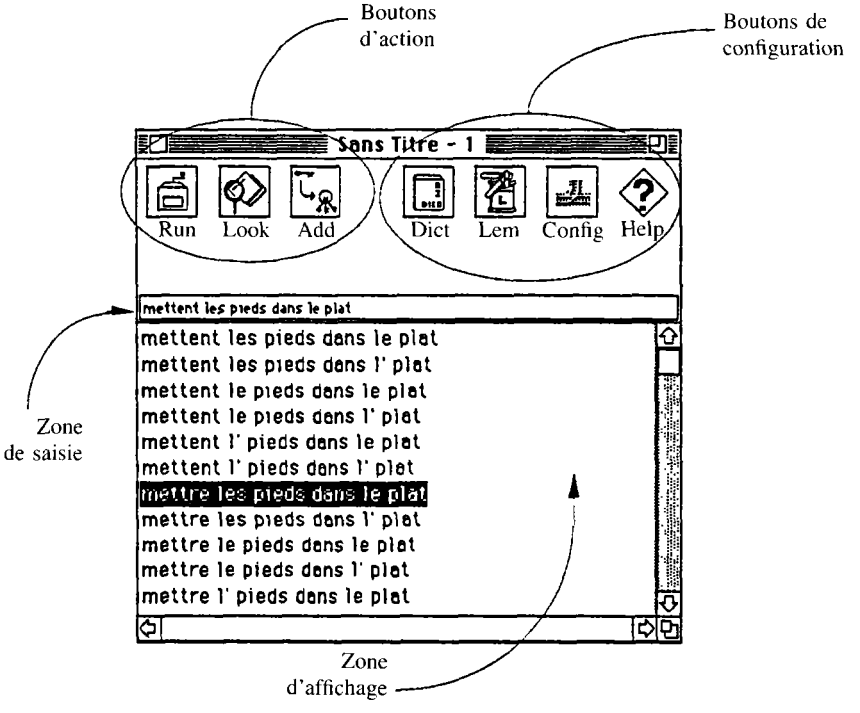


Figure 1 : La fenêtre principale d'Odile. Cette fenêtre est divisée en deux parties : son bord (partie supérieure) et son contenu (partie inférieure). Le bord comporte deux séries de boutons. La première correspond à des boutons performant des actions, la seconde permet de configurer les fonctionnalités d'Odile. Le contenu est une zone de texte libre ou sont affichées les clés obtenues par combinaison des formes et des lemmes.

Il sélectionne ce fragment, le copie et lance Odile. La fenêtre d'Odile s'affiche. L'utilisateur colle alors le fragment dans la *zone de saisie* d'Odile (voir figure 1) et lance la génération de clés avec le bouton *Run*.Les clés (potentielles) calculées s'affichent alors dans la *zone d'affichage*. On peut automatiser le processus en configurant Odile différemment. Les différentes possibilités sont illustrées plus bas avec l'automate de contrôle.

L'utilisateur sélectionne la clé désirée et lance la recherche (bouton *Look*) dans le dictionnaire. La recherche se fait par défaut dans le dictionnaire usuel, mais on peut aussi chercher dans d'autres dictionnaires.

Si les clés proposées ne lui conviennent pas, l'utilisateur a toujours la possibilité d'en entrer une directement au clavier. Il peut aussi ajouter une clé dans son « lexique personnel ». Il peut bien sûr spécifier le ou les lexiques personnels où se fera la recherche.

Une fois la recherche terminée, la fenêtre de l'outil dictionnaire (WinTool™) s'affiche. Si la clé a été trouvée, on y voit l'article correspondant (clé + contenu). Si la clé était absente du dictionnaire, la fenêtre contient l'article de clé immédiatement inférieure.

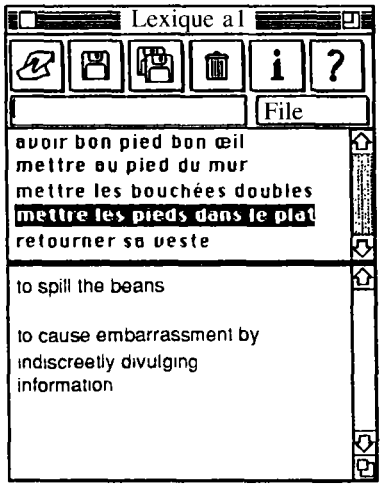


Figure 2 : Une fiche WinTool coorespondant à l'expression « mettre les pieds dans le plat ».

Les icônes de l'outil dictionnaire permettent la gestion de la base ainsi que la navigation. Une fois les informations trouvées, l'utilisateur retourne à son traitement de texte. Il peut garder la fenêtre affichée à côté de la fenêtre contenant son article de manière à avoir toujours accès à ce qu'il a cherché. Il peut également copier et coller les informations.

L'automate de contrôle

La séquence de traitements effectuée à partir de l'introduction d'un fragment dans la zone d'édition et de l'appui du bouton *Run* peut s'illustrer par un automate (figure 3).

Cet automate permet de définir le traitement à appliquer en fonction des options choisies par l'utilisateur. Pour le moment, deux options (majeures) sont définies :

- vérifier d'abord si la chaîne rentrée constitue une entrée de lexique ou de dictionnaire. Si oui, il n'y a pas de lemmatisation et Odile affiche directement l'article correspondant;
- utiliser un lexique personnel pour produire directement des clés de recherche.

Dans ce cas, les clés potentielles sont utilisées pour accéder à un lexique personnel (« trousseau »). Le contenu de ces clés est ensuite « interprété » comme une liste de clés, qui sont alors utilisées pour chercher dans le dictionnaire.

Dans l'automate, ces deux options correspondent respectivement aux conditions C1 et C2.

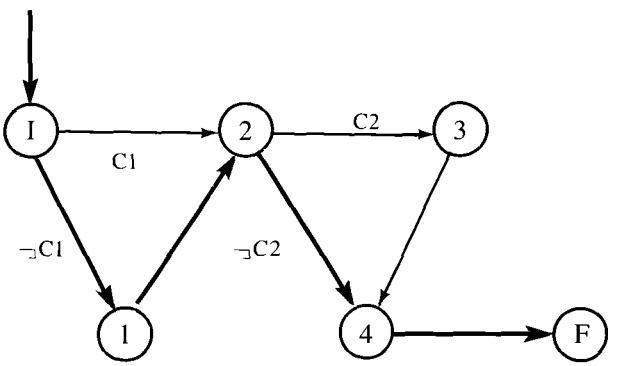


Figure 3 : L'automate de contrôle d'Odile.
Les états I et F sont respectivement les états initial et final. Les Ci sont les conditions pour passer d'un état à l'autre. Les arcs vides sont des transitions par lesquelles on passe sans conditions. Le chemin en gras correspond au plus court chemin possible, celui où l'utilisateur n'utilise aucune des options de traitement.

L'automate a six états. I est l'état initial, F l'état final;

- état 1 : on lemmatise F (la forme d'entrée). On obtient [CP]+, une liste non vide de clés potentielles CP;
- état 2 : état vide, on ne fait rien (c'est un état de transition);
- état 3 : on effectue la recherche des [CP]+ dans le(s) lexique(s) personnel(s) courant(s). [CP]+ devient alors la liste des clés trouvées comme contenu des entrées du(des) lexique(s) personnel(s);

– état 4 : on effectue une recherche dans la(les) base(s) dictionnaires des entrées [CP]+ et on affiche la première clé trouvée. On affiche également dans la fenêtre Odile la liste des clés potentielles.

Standardisation des lemmatiseurs (définition d'une interface universelle pour les lemmatiseurs)

Nous avons élaboré des points d'entrée standard et des structures de sortie génériques pour tous les lemmatiseurs possibles. Les structures de données d'entrée sont décrites avec les points d'entrée.

Structure de données d'entrée pour les lemmatiseurs

Une chaîne constitue les données d'entrée. Elle a la forme <chaîne sans espace> (espace <chaîne sans espace>)*. Une telle chaîne est dite normalisée.

Structure de données de sortie pour les lemmatiseurs

La structure de sortie est un graphe avant contenant les informations suivantes :

– les nœuds représentent des formes ou des lemmes;

– les arcs représentent les chemins possibles parmi ces nœuds;

– chaque nœud contient jusqu'à dix informations (entre parenthèses sont indiqués le nom de la variable et son type) :

1) le numéro du nœud (Numéro : entier);
2) son identification, obligatoire (valChaîne : chaîne) et ne contenant pas de blanc;
3) son type, obligatoire (TypeChaîne) : Forme ou Lemme (F | L);
4) sa nature, facultative (Nature) : Occurrence complète, Tournure, morceau Gauche, Interne, ou Droit (D | O | I | G | T);
5) sa localisation, obligatoire (Origine : entier), désigne l'origine (0 désignant la position précédant le premier caractère) de ValChaîne;
6) sa longueur, obligatoire (Longueur : entier), qui est le nombre de caractères couverts par le nœud (la longueur de la sous-chaîne traitée dans la chaîne d'entrée);
7) sa catégorie morphologique (Catégorie : chaîne), facultative (par exemple : (Catégorie 'adv) ou (Catégorie 'subc) ou encore (Catégorie 'dept));
8) ses variables morphologiques (Variables : liste), facultatives (par exemple : (Variables('plus 'mas 'tre 'suj)) ou (Variables((nombre 'plus) (genre 'mas) (fonction 'suj)));

9) des informations supplémentaires (Reste : liste), facultatives, qui pourraient être utilisées dans l'avenir (par exemple, relief typographique, domaine du document, etc.);

10) les numéros des nœuds successeurs du nœud courant dans le graphe (Suivant : liste d'entiers), la liste vide (()) dénotant l'absence de successeur.

Cette structure est représentée dans un fichier texte, qui contient la suite des nœuds dans l'ordre ascendant des numéros. Chacun des nœuds est codé (à la LISP) comme suit :

(((Numéro *Numéro*) (ValChaîne *ValChaîne*)
(TypeChaîne *TypeChaîne*) [(Nature *Nature*)]
(Origine *Origine*) (Longueur *Longueur*)
[(Catégorie *Catégorie*)] [(Variables ({*Variables*}+))]
[(Reste ({*Reste*}+))])
(Suivants ({*Suivant*}+)

Notation : les caractères spéciaux { } [ ] * + sont utilisés pour coder des expressions régulières. Les parties entre crochets ([ ]) sont facultatives, les symboles en italiques dénotent les valeurs et les accolades ({ }) permettent des regroupements. La construction {*x*}+ signifie que x peut être répété une ou plusieurs fois, et {*x*}* signifie que x peut être répété zéro ou plusieurs fois.

Points d'entrée pour les lemmatiseurs

Procédure Lem-ChargerDonnees;

La procédure Lem-ChargerDonnees charge les données nécessaires à l'analyse des occurrences fournies en entrée.
L'appel de cette procédure doit se faire une seule fois (avant de lancer la procédure Lem-Lemmatiser!).

Procédure Lem-Lemmatiser (Chaîne : Chaîne;
var StructSortieLem : Fichier texte);

La procédure Lem-Lemmatiser lemmatise le groupe d'occurrences *(Chaîne)* fourni en entrée. Elle construit la structure des résultats de lemmatisation *StructSortieLem*.

Chaîne est la chaîne de caractères à lemmatiser. Il s'agit d'une chaîne « normalisée », décrite par l'expression régulière : <chaîne sans blanc>(blanc <chaîne sans blanc>)*.

StructSortieLem est la structure de sortie, construite dans tous les cas.

Exemple

Avec un lemmatiseur approprié, l'analyse de la chaîne « les poules couvent » peut donner le fichier texte résultat suivant. Les informations optionnelles Catégorie et Variables ne figurent pas.

(((Numéro 1) (ValChaîne 'les) (TypeChaîne 'F) (Nature 'O) (Origine 1) (Longueur 3)) (Suivants (4 5)))

(((Numéro 2) (ValChaîne 'le) (TypeChaîne 'L) (Nature 'O) (Origine 1) (Longueur 3)) (Suivants (4 5)))

(((Numéro 3) (ValChaîne 'il) (TypeChaîne 'L) (Nature 'O) (Origine 1) (Longueur 3)) (Suivants (4 5)))

(((Numéro 4) (ValChaîne 'poules) (TypeChaîne 'F) (Nature 'O) (Origine 5) (Longueur 6)) (Suivants (6 7 8 9)))

(((Numéro 5) (ValChaîne 'poule) (TypeChaîne 'L) (Nature 'O) (Origine 5) (Longueur 6)) (Suivants (6 7 8 9)))

(((Numéro 6) (ValChaîne 'couvent) (TypeChaîne 'F) (Nature 'O) (Origine 11) (Longueur 7)) (Suivants ()))

(((Numéro 7) (ValChaîne 'couvent) (TypeChaîne 'L) (Nature 'O) (Origine 11) (Longueur 7)) (Suivants ()))

(((Numéro 8) (ValChaîne 'couver) (TypeChaîne 'L) (Nature 'O) (Origine 11) (Longueur 7)) (Suivants ()))

(((Numéro 9) (ValChaîne 'couver) (TypeChaîne 'L) (Nature 'O) (Origine 11) (Longueur 7)) (Suivants ()))

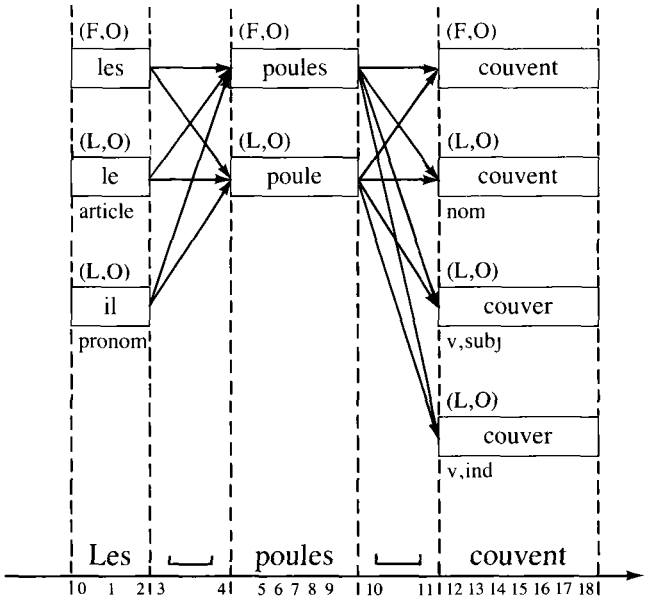


Figure 4 : Résultat de la lemmatisation de la chaîne (normalisée) « les poules couvent ».

Intégration d'un outil dictionnaire

Pour qu'on puisse l'intégrer à Odile, un outil dictionnaire (et WinTool) doit avoir les points d'entrée suivants :

Chargement et déchargement de l'outil dictionnaire

Fonction LoadDT : Erreur;

Cette procédure permet de charger les données nécessaires pour le fonctionnement de l'outil dictionnaire

Fonction UnloadDT : Erreur;

Cette procédure doit être appelée quand on ne veut plus utiliser l'outil dictionnaire.

Contrôle de l'outil dictionnaire

Procédure ControlDT (Event : Événement);

On passe l'événement *(Event)* à l'outil dictionnaire.

Gestion des bases

Fonction OpenDTBase (BaseName : chaîne;
InfoNeeded : data;
Lexique : Booléen) : Erreur;

Cette fonction ouvre une base de l'outil dictionnaire.

Fonction CloseDTBase (BaseIndex : entier) : Erreur;

Cette fonction ferme une base de l'outil dictionnaire.

Gestion des articles

Fonction FindArticle (Key : string;
BaseIndex : integer;
Seuil : integer) : Erreur;

Cette procédure permet de rechercher des articles dont l'entrée est proche de *Key* dans la base *BaseIndex*. Proche signifie que la distance entre les articles trouvés et *Key* est inférieure à *Seuil*. Cette fonction écrit dans un fichier texte une liste de doublets. Chaque doublet représente la clé de l'article trouvé et la distance de cet article à *Key*. Le format du fichier est :

[(c,d)]* ()

c est une clé d'article et *d* est la distance de c par rapport à *Key*. Une liste vide (*()*) indique la fin de la liste.

On utilise cette procédure quant on veut juste s'assurer de l'existence d'un article dans une base, mais que l'on ne veut pas récupérer l'information associée.

Fonction GetArticle (Key : chaîne;
KeyNumber : integer;
BaseIndex : entier;
var Content : liste de données;
Visible : boolean) : Erreur;

Cette procédure permet d'obtenir les informations associées à l'article repéré par la clé *Key* de la base repérée par *BaseIndex*.

S'il existe plusieurs articles ayant la même clé, on obtiendra les *KeyNumber* premiers. On utilisera cette fonction pour récupérer les informations d'un article dont on sait qu'il existe dans la base.

Fonction PutArticle (Key : string;
Content : data;
BaseIndex : integer) : Erreur;

Cette procédure crée un nouvel article de dictionnaire dans la base référencée par *BaseIndex*. Cet article a pour clé *Key* et pour contenu les informations (Texte) référencées par *Content*.

Perspectives

Nous travaillons actuellement sur un certain nombre d'améliorations par rapport à la version actuelle.

Accélération du traitement

Le lemmatiseur d'origine (PILAF [4] et [3]) n'est pas vraiment adapté à ce que nous voulons faire. La structure de dictionnaire couplée au lemmatiseur n'est pas très efficace. Elle peut être améliorée, diminuant ainsi grandement le temps de génération de clés.

Nous cherchons également à adapter un ou plusieurs autres lemmatiseurs, puisqu'Odile a été conçu pour offrir une telle modularité.

Amélioration de l'interface utilisateur

L'idée d'une fenêtre « vocable » (d'après la revue du même nom) enfin possible avec le système 7.0 d'Apple [2] serait très intéressante. Une telle fenêtre serait disposée à côté d'une fenêtre de travail et contiendrait les articles des dictionnaires usuels correspondant aux termes présents dans la fenêtre. Elle serait effectivement liée à la fenêtre de travail en ce sens que l'outil irait « par avance » rechercher les informations associées au vocabulaire du texte affiché.

Conclusion

L'augmentation des performances des micro-ordinateurs permet maintenant la conception d'outils grand public en THAM et dans bien d'autres contextes. Les outils plus spécifiques s'intégreront à des environnements spécialisés pour traducteurs professionnels.

Une version d'Odile existe déjà. La version suivante, en préparation, sera une refonte complète tournant sous la version 7 du système d'exploitation du Macintosh, et organisée autour d'applications communicantes. Elle devrait permettre la réalisation de fenêtres « vocables », ce qui semble impossible ou extrêmement délicat avec les versions précédentes du système.

En effet, la réalisation technique d'outils dépend beaucoup des possibilités de l'environnement en général, et du système d'exploitation en particulier. Certains systèmes d'avant-garde offrent de bonnes idées. Par exemple, la Communication Inter-Application de la version 7.0 du Macintosh doit permetre un prototypage poussé des fenêtres vocables en attendant la généralisation de ce concept à d'autres environnements. L'émergence de standards ou de systèmes d'exploitation équivalents comme UNIX avec XWindow/Motif et OS2/PM en permettra sans doute la large diffusion.

La démarche adoptée consiste à développer les idées nouvelles sur le premier système disponible permettant leur réalisation, tout en prévoyant une adaptation ultérieure pour les autres environnements fonctionnellement équivalents qui ne manqueront pas de voir le jour.

Bibliographie

[1] APPLE Computer Inc (1985) *Inside Macintosh*, Volume I, Addison-Wesley Publishing Company, Inc., 550 p.

[2] APPLE Computer Inc (1990) *Inside Macintosh*, Volume V, Addison-Wesley Publishing Company, Inc., 1200 p.

[3] CHAPON L., DJAMEI M., DJOHARIAN P., MORENO F. (1989) *Étude et révision du système PILAF, Portage vers le Macintosh, Réalisation d'un lemmatiseur, Rapport final de projet,* Projet DESS-IDC, juin 1989.

[4] COURTIN J., DUJARDIN D., GENTHIAL D., KOWARSKI I., COHARD B., STRUBE DE LIMA V. « Le système PILAF », document interne LGI (IMAG) présenté au *PRC Communication Homme-Machine*, février 1988.

[5] COUTAZ J. (1988) *Interface Homme-Ordinateur : conception et réalisation*, thèse de doctorat d'État (UJF), décembre 1988, 402 p.

[6] HUNDT M. G. (1982) *Working with the Weidner Machine-Aided Translation System, Practical Experience of Machine Translation*, V. Lawson (éd.), North-Holland Publishing Compagny, ASLIB, p. 45-51.

[7] KJAERSGAARD P. S. (1987) « REFTEX — Un progiciel pour la traduction assistée par ordinateur », *Computer and Translation*, 7 p.

[8] LONSDALE D. (1984) « Notes on Interactive Translation », *Automated Language Processing Systems (ALPS)*, Provo, Utah, 10 p.

[9] MELBY Alan K. (1978) « Design and Implementation of a Computer-Assisted Translation System », Présenté à la *7th International Conference on Computational Linguistics (COLING-78)*, Bergen, Norvège, 14-18 août 1978, 28 p.

[10] MELBY Alan K. (1984) *MERCURY™ Card File and Glossary Manager, User's Guide,* LinguaTech, Provo, 114 p.

[11] META, (1990) *Journal des traducteurs*, 35e anniversaire, Les Presses Universitaires de Montréal, Vol. 35, n° 1, mars 1990.

[12] PALMER P. (1990) *Étude d'un analyseur de surface de la langue naturelle, Application à l'indexation automatique de textes,* thèse (UJF), septembre 1990, 292 p.

[13] TOMASINO I. (1990) *ODILE : un outil d'intégration extensible de dictionnaires et de lemmatiseurs*, mémoire (CNAM), décembre 1990, 151 p.

[14] TONG L. C. (1986), *SISKEP — A Software for the Human Translator*, Computer Aided Translation Project, Université Sains Malaysia, Penang, 6 p.

[15] TONG L. C. (1987) « The Engineering of a Translator Workstation », *Computers and Translation 2*, 11 p.

[16] WinSoft (1987) *Manuel d'utilisation de WinTool™, Version 1.1* WinSoft.

10

Intégration des connaissances en génie linguistique : niveaux, dimensions, objets et contraintes*

Jean-Yves MORIN

Université de Montréal, Montréal, Canada

En ce qui concerne leur langage,
les honnêtes gens (君子 *jūnzǐ*)
ne laissent rien au hasard.
Confucius, *Entretiens familiers*

Introduction

Problème fondamental du génie linguistique

Le problème fondamental du génie linguistique est l'élaboration d'environnements de travail pour les linguistes, les traducteurs et les rédacteurs permettant la manipulation intelligente d'objets sémiotiques de type linguistique[1].

* La recherche sous-jacente à ce travail a été rendue possible grâce à des subventions du Conseil de recherches en sciences naturelles et en génie du Canada, de l'Office de la langue française du Québec et de la Coopération France-Québec (ministère des Affaires internationales, Québec et ministère des Affaires étrangères, France) ainsi que grâce à des contrats de recherche avec le Centre canadien de recherches en informatisation du travail (CCRIT, ministère des Communications, Canada). Les opinions exprimées ici n'engagent évidemment que l'auteur.

1. Les objets sémiotiques de type graphique ou musical ne relèvent donc pas du génie linguistique, pas plus que les textes structurés selon des systèmes textuels articulés qui ne sont pas des langues naturelles (programmes, expressions mathématiques, formules chimiques, etc.).

De tels environnements devront disposer d'une quantité importante de connaissances linguistiques. Ces connaissances devront éventuellement être combinées à des connaissances extra-linguistiques, de nature encyclopédique ou pédagogique, par exemple[2].

Le volume et la disparité de ces connaissances linguistiques pose de délicats problèmes d'INTÉGRATION[3]. D'une part, ces connaissances sont traditionnellement partitionnées en NIVEAUX plus ou moins autonomes qu'il est essentiel de pouvoir interfacer les uns aux autres de façon efficace[4]. D'autre part, même à l'intérieur d'un niveau les REPRÉSENTATIONS utilisées pour caractériser les connaissances linguistiques peuvent être très différentes.

Ainsi, pour caractériser l'*ordre des constituants* à l'intérieur d'un syntagme, une théorie comme la théorie gouvernement-liage (Chomsky (1981), dorénavant GL) utilisera essentiellement trois paramètres binaires (tête initiale/finale, assignation de Cas vers la gauche/droite et assignation de θ-rôle vers la gauche/droite)[5], tandis qu'une

---

Nous ne nous intéresserons ici qu'aux problèmes liés à la manipulation intelligente de « textes » au sens classique, c'est-à-dire d'objets structurés autour d'une représentation linéaire. C'est dire que nous ne tiendrons pas compte des développements en ce qui concerne les HYPERTEXTES structurés autour de représentations hiérarchiques ou hétérarchiques. Il est évident que les contraintes apportées par des systèmes dotés d'HYPER-MODÈLES ou d'HYPER-CRITIQUES changeront certainement de façon radicale la nature des activités de rédaction et de traduction dans les domaines où elles seront applicables. Voir à ce sujet Barrett, E. (1988) et (1989 réd.), par exemple.

Nous n'avons pas non plus tenu compte de la dimension stylistique et esthétique.

2. À titre d'exemple, mentionnons que les « connaissances linguistiques » intégrées aux traitements de texte universels actuels sont limitées à la détermination de segments linéaires : caractères, mots, phrases, paragraphes, pages, sections, etc. Un segment donné ne diffère d'un autre que par sa longueur, sa position et la séquence des marques qui le composent.

   Notons cependant qu'un bon traitement de texte dispose également de connaissances typographiques sur les en-têtes, les pieds de page, les marges, les alinéas, les espacements, l'orthographe (non grammaticale), la césure, la justification, les renvois et les notes infra-paginales, les titres, la hiérarchie des parties d'un texte, les types de caractères (polices, tailles, styles, corps), etc.

   Il faut distinguer les systèmes de traitement de texte universels éditeurs de structures (tels EMACS), qui, eux, sont dotés de certaines connaissances supplémentaires sur les types de structures qu'ils manipulent (programmes LISP dans le cas d'EMACS). La plupart des traitements de texte seront bientôt dotés de certaines connaissances structurales sur des types de textes standards (rapports, lettres, bulletins d'information, articles, bibliographies, etc.) et sur leurs divisions habituelles. Mais, ici encore, on touche à l'univers de l'hypertexte, plutôt qu'à celui de la langue naturelle.

3. Sans compter les problèmes D'ÉLABORATION. En effet, les connaissances explicites sur la nature et le fonctionnement du langage sont encore assez rudimentaires. Malgré certaines déclarations fracassantes, les progrès réalisés au niveau de la description des langues particulières et de la compréhension des mécanismes généraux du langage au cours des trente dernières années sont loin d'avoir été foudroyants.

4. Ce problème ne se pose généralement pas avec une telle acuité au linguiste, puisque ces différents niveaux correspondent plus ou moins aux sous-disciplines de la linguistique. Les descriptions se situent le plus souvent à l'intérieur d'un seul niveau. Voir cependant Zwicky (1986) pour un ensemble de travaux récents sur ce problème.

5. *Cf.* Travis (1989) et Dean 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 GL, puisque cela ajoute à la « richesse de la structure déductive ». Du point de vue de la représentation des connaissances linguistiques, c'est plutôt là un défaut, puisque cette propriété empêche de maintenir la transparence fonctionnelle des descriptions et donc la modularité des représentations.

théorie comme la grammaire syntagmatique généralisée (Gazdar *et al.* (1985), dorénavant GSG) utilisera un type particulier de règles, les PL-règles.

Par exemple, en français, les compléments apparaissent après la tête lexicale (verbe, nom, adjectif ou préposition) dont ils dépendent, alors qu'en japonais, ils apparaissent avant.

(1) (a) *Pinocchio frappe le chemin avec ses pieds de bois*[6]

(b) [P [SN *Pinocchio* SN]
[SV [V *frappe* V]
[SN *le chemin* SN]
[SP [Prép *avec* Prép]
[SN [Dét *ses* Dét]
[N *pieds* N]
[SX *de bois* SX]
SN]
SP]
SV]
P]

6. Pour des raisons de simplicité, nous avons choisi une représentation par crochets étiquetés indentée de la structure syntagmatique. Cette représentation utilise des étiquettes qui sont des abréviations des catégories utilisées en GSG plutôt que de celles utilisées en GL. Structuralement, elle correspond essentiellement à une représentation GSG.

Une représentation de la structure syntagmatique en GL serait nettement plus complexe que celle ci-dessus, mettant en jeu plusieurs niveaux intermédiaires (SC, SI, SD) et plusieurs catégories vides (Infl, Comp, Dét, etc.), mais, pour ce qui nous concerne ici (la distribution des catégories lexicales effectivement réalisées relativement aux catégories syntagmatiques), elle serait essentiellement congruente à la représentation GSG.

La représentation de la structure syntagmatique de (1) en GL serait, en simplifiant quelque peu, quelque chose comme (i).

(i) [SC [Comp Comp]
[SI [SD [Dét Dét]
[SN *Pinocchio* SN]
SD]
[Infl Infl]
[SV [V *frappe* V]
[SN *le chemin* SN]
[SP [Prép *avec* Prép]
[SD [Dét *ses* Dét]
[SN [N *pieds* N]
[SX *de bois* SX]
SN]
SD]
SP]
SV]
SI]
SC]

(2) (a) *Pinokkio-wa ki no ashi de michi-o keri.*
'*Pinocchio*-WA *bois*-NO *pied*-DE *chemin*-O *frapper*'

(b) [P [SN *Pinokkio-wa* SN]
[SV [SP
[SN
[SP *ki no* SP]
[N *ashi* N]
SN]
[Postp *de* Postp]
SP]
[SN *michi-o*]
[V *keri* V]
SV]
P]

En GL, on posera donc que le français est une langue 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[7].

(3) Paramètres GL

(a) Français

Tête	Cas	Θ-rôle
Initiale	**Droite**	**Droite**

(b) Japonais

Tête	Cas	Θ-rôle
Finale	**Gauche**	**Gauche**

7. En fait, la description des contraintes d'ordre des mots en français (et en japonais) dans le cadre GL est beaucoup plus complexe que ne le laissent soupçonner ces tableaux. La valeur des trois paramètres est 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, certaines têtes doivent être initiales, d'autres finales. On note par ailleurs que ce système contraint à une multiplication de niveaux, puisqu'il exclut *a priori* des structures « plates » où la tête est enrobée de part et d'autre de constituants (spécificateurs d'un côté et compléments de l'autre, par exemple [SN Dét N SP] ou [SV Aux V SN] en français), sauf dans une langue où le paramètre de tête (initiale/finale) ne serait pas actif.

Dans le cadre GSG, on posera que ce sont les CATÉGORIES LEXICALES (têtes, mais également spécificateurs — déterminants, auxiliaires, modificateurs, etc.) qui doivent apparaître avant les catégories syntagmatiques en français et après ces dernières en japonais.

(4) (a) Français

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

(b) Japonais

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

Si l'on tentait de calquer la théorie GL en GSG, on pourrait postuler un trait [TÊTE : α] caractérisant les têtes de syntagmes[8] et avoir des PL-règles comme en (5).

(5) (a) Français

PL : [Tête : α] ≪ ¬[Tête : α]

(b) Japonais

PL : ¬[Tête : α] ≪ [Tête : α]

On voit tout de suite la difficulté à intégrer une description du SN (français ou du japonais) suivant le cadre GSG à une description de la proposition conforme à GL, par exemple.

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[9].

8. 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 selon le NIVEAU (lexical ou non lexical) plutôt que la fonction (tête *vs* spécificateurs ou compléments).
9. Quant au problème de l'intégration des NIVEAUX, il demeure, de toute façon, incontournable.

L'expérience montre que tel n'est pas le cas[10]. 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?

Représentation des connaissances linguistiques

Objets et contraintes

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.

Objets

Représentation à base d'objets : grammaires à base lexicale

Plusieurs théories grammaticales récentes[11] 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

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

11. Par exemple GL, où lexique et paramètres sont fondamentaux, la grammaire lexicale-fonctionnelle (dorénavant GLF) ou la grammaire syntagmatique endocentrique (*head-driven phrase structure grammar*, dorénavant GSE), Pollard (1984), Sag & Pollard (1987).

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[12].

Dans les théories linguistiques basées sur la notion d'*information grammaticale*[13], 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[14].

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 ce qui concerne les ressources. Les autres objets (structures, relations) ne sont accessibles que par construction[15] et sont donc de coût variable (éventuellement prohibitif) toujours en ce qui a trait aux 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 (par exemple, catégories, catégories lexicales, catégories lexicales majeures ou verbes, verbes transitifs, verbes transitifs à double objet, etc.). 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 stockée dans la classe de l'objet (par exemple, du verbe) et n'a pas à être répétée 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.

Ce type de représentation permet de construire des descriptions de façon très modulaire, incrémentale et dynamique et d'exprimer des valeurs par défaut.

On peut aller jusqu'à poser l'hypothèse que toute la connaissance grammaticale peut être représentée en objets.

Non seulement les objets lexicaux sont associés à des règles (ou des familles de règles[16]), mais les règles elles-mêmes peuvent être vues comme des objets abstraits.

---

12. Notons cependant les travaux de Gross (1975) sur les *lexiques-grammaires* et ceux de Mel'čuk (1984, 1988) sur les *dictionnaires explicatifs et combinatoires* (DEC) qui, malgré leur intérêt, s'intègrent difficilement aux théories syntaxiques issues de la grammaire générative.
13. Sur cette notion, *cf.* Sag & Pollard (1987).
14. Sur le problème de la couverture, *cf.* Bouchard, Emirkanian et Morin (1991).
15. 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.
16. *Cf.* Morin (1989), Blache (1990), Blache & Morin (1990) sur différentes notions de *famille de règles*.

Les règles et les principes sont organisés en réseaux de classes et de sous-classes (correspondant en partie aux « modules » d'une théorie comme GL ou GSG). Elles ont des propriétés qui peuvent être héritées (ou déléguées).

Ainsi, une règle ou un principe en GL 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 GL peut être vue comme une classe d'objets ayant des paramètres de CAS (eux-mêmes formés d'une VALEUR et d'un ASSIGNATEUR), d'une DIRECTION et d'une STRATE d'application (qui serait, par défaut la *S-tructure*).

(6) Assignation de Cas (GL)

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

De même, en GSG, une DI-règle 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.

(7) DI-règle (GSG)

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

Les GRAMMAIRES D'OBJETS constituent donc un métaformalisme où différents types de connaissances grammaticales (correspondant, par exemple, à différentes théories linguistiques) peuvent être exprimés.

L'élaboration de (méta)formalismes 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 GL)[17].

En fait, les grammaires de principes 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[18].

Ainsi, en formulant une notion comme celle de *c-commande* en GL, on n'a pas à se soucier de la façon précise dont seront représentées les structures syntaxiques ni des étiquettes sur les nœuds.

(8)

C-COMMANDE

α *c-commande* β si et seulement si α et β ne se dominent pas l'un l'autre et que le premier nœud branchant qui domine α domine également β.

Dans cette définition, α et β sont implicitement typés comme des nœuds (étiquetés) dans une structure arborescente σ, également implicite.

17. En fait, cette opposition est plus méthodologique que 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).
18. D'ailleurs, toute contrainte constitue elle-même une fonction de typage abstrait.

On a donc la définition suivante :

(9)

C-COMMANDE

c-commande (α, β, σ) **ssi** (*PNBD* (γ, α, σ)
$\supset$
domine (γ, β, σ))

Cette définition présuppose celle d'un prédicat *domine*, dont les trois arguments (α, β et σ) sont des nœuds étiquetés (α et β) et une structure (σ) et qui lui-même présuppose un prédicat *sous-arbre*, qui peut être défini à partir de primitifs d'égalité, d'appartenance et de représentation d'arbres[19] :

(10)

DOMINE

domine (α, β, σ) **ssi** *sous-arbre* ($\alpha(\lambda)$, σ)
et
sous-arbre ($\beta(\lambda')$, $\alpha(\lambda)$)
et
$\neg$*égal* ($\alpha(\lambda)$, $\beta(\lambda')$)

(11)

SOUS-ARBRE

sous-arbre (σ, $\alpha(\lambda)$) **ssi** (*égal* (σ, $\alpha(\lambda)$)
ou
(*membre* (σ', λ)
et
sous-arbre (σ, σ')))

19. Dans cette représentation, le terme $\alpha(\lambda)$ est une abréviation pour un arbre étiqueté dont α est la racine étiquetée et dont λ est la liste (éventuellement vide) des sous-arbres immédiats.

Les variables à gauche du connecteur **ssi** sont quantifiées universellement et les variables non liées à sa droite sont quantifiées existentiellement (*i.e.* de façon constructive).

Les prédicats *branchant* et *premier nœud branchant dominant* (*PNBD*) sont également implicites.

(12)

BRANCHANT

branchant (α, σ) **ssi** *sous-arbre* ($\alpha(\lambda)$, σ)
et
longueur (λ, 1)[20]
et
$1 > 1$

(13)

PNBD

PNBD (α, β, σ) **ssi** *domine* (α, β, σ)
et
branchant (α, σ)
et
(*domine* (γ, α, σ)
ou
$\neg$ (*domine* (γ, β, σ)
et
branchant (γ, σ)))

On peut éliminer cette définition un peu lourde en tenant compte de la transitivité de la relation de dominance, qui entraîne que si le premier nœud branchant qui domine α domine également β, alors *tous* les nœuds branchants dominant α dominent β.

Dès lors on n'a plus besoin, pour définir *c-commande*, du prédicat *PNBD*, mais uniquement des prédicats *domine* et *branchant*.

(14)

C-COMMANDE

c-commande (α, β, σ) **ssi** ((*branchant* (γ, σ)
et
domine (γ, α, σ))
$\supset$
domine (γ, β, σ))

20. Le prédicat *longueur* relie un premier argument de type *séquence* à un deuxième argument de type *nombre entier* représentant la longueur de cette séquence.

C'est donc sur le domaine (infini) des arbres étiquetés que porte la contrainte.Cette définition est purement relationnelle et totalement indépendante de la nature et de la représentation précise des objets qu'elle concerne.

On retrouve les mêmes possibilités d'abstraction sur les données dans des approches ascendantes, basées sur des règles générales. Par exemple, en GSG, une PL-règle comme (7) est tout à fait indépendante des représentations choisies, tant pour les structures catégorielles que pour les arbres.

(15)

[Spéc : α] $\ll$ α[21]

Contraintes et domaines

Le problème fondamental de l'intégration des descriptions linguistiques est qu'il n'existe pas de métalangage commun aux différentes théories linguistiques actuelles. Il est donc difficile de pouvoir évaluer exactement quelles sont les contraintes qu'elles apportent sur ce que Milner (1989) appelle le « possible de langue », c'est-à-dire comment elles limitent effectivement la notion de « langue naturelle possible ».

Trop souvent, on confond la notion de contraintes sur les « langues naturelles possibles » et celle, beaucoup moins intéressante, de contraintes sur les « descriptions possibles » de langues naturelles. Si l'on contraint les descriptions des langues naturelles à être exprimées dans un système n'utilisant que l'alphabet romain, par exemple, excluant ainsi les caractères grecs, ou à n'être rédigées qu'en latin classique, on apporte

---

21. *Cf.* Morin (1989).

Ici α subsume MAJEUR : β / NIVEAU : γ / FORME : δ et β, γ et δ sont dans la portée de MAJEUR, NIVEAU et FORME respectivement (*i.e.* $\beta \in \rho$(MAJEUR), $\gamma \in \rho$(NIVEAU) et $\beta \in \rho$(FORME)) la portée (ρ) est définie pour tous les traits et non seulement pour les traits à valeur atomique.

Le prédicat théorique PL, correspondant à l'opérateur $\ll$ peut être défini de la façon suivante :

PL (α, β, $\gamma(\lambda)$) **ssi** $\neg$*précède* (β, α, λ)

Le prédicat *précède* peut à son tour être défini récursivement en termes des primitifs *égal* et *membre*.

précède (α, β, λ) **ssi** (*égal* (λ, $\alpha(\xi).\lambda'$) **et** *membre* ($\beta(\chi)$, λ')) **ou** (*égal* (λ, $\gamma(\varphi).\lambda'$) **et** $\neg$*égal* (β, γ) **et** *précède* (α, β, λ')))

Le point dans $\alpha(\xi).\lambda'$ et $\gamma(\varphi).\lambda'$ est l'opérateur de concaténation de séquence (le CONS de liste). Si la séquence $\lambda = \alpha.\lambda'$, α est le premier élément de la séquence λ et λ' le « reste » de cette séquence.

des contraintes, très fortes, sur la forme des descriptions possibles, mais absolument vides quant au contenu desdites descriptions.

Les contraintes sur la forme et le fonctionnement des grammaires n'ont de sens que si elles ont des conséquences directement déductibles sur la classe des objets (langages) décrits, sinon il ne s'agit que de variantes notationnelles. Ainsi, pour prendre un autre exemple, un peu plus technique, le fait de poser des contraintes très fortes sur les règles d'une grammaire indépendante du contexte telles que :

(16) (a) les règles sont de la forme (forme normale de Chomsky, FNC)

$A \longrightarrow B\ C$ ou $A \longrightarrow a$

ou encore

(b) les règles sont de la forme (forme normale de Greibach, FNG)

$A \longrightarrow a\ \varphi$ ou $A \longrightarrow a$

où A, B et $C \in V_n$ (le vocabulaire non terminal de la grammaire),
$a \in V_t$ (le vocabulaire terminal de la grammaire) et
$\varphi \in (V_n \cup V_t)^*$

ne change en rien la classe des objets que l'on peut décrire, même si cela réduit radicalement la classe des grammaires possibles. Pour ce qui est des objets qui peuvent être décrits, il s'agit toujours de la classe tout entière des langages de type 2. Si la réduction d'une grammaire à l'une ou l'autre de ces formes normales peut avoir un intérêt (pour l'écriture d'algorithmes de reconnaissance d'un certain type, ascendant pour FNC, prédictif pour FNG, par exemple), ce n'est certainement pas un intérêt théorique.

Comment comparer, par exemple, les hypothèses sur le liage[22] d'une théorie comme gouvernement-liage, et celles d'une théorie comme la grammaire syntagmatique généralisée[23] ? Il faudrait pouvoir définir les théories en fonction des *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 (1989)).

En traitement des langues naturelles, cette nécessité théorique s'associe à des exigences pratiques. Idéalement, on voudrait pouvoir pratiquer un certain éclectisme

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

éclairé, qui permettrait d'importer dans un système $\mathbb{S}$ les contraintes exprimées par la théorie $\mathbb{A}$ sur un domaine $\mathbb{D}$, de les combiner avec les contraintes exprimées par la théorie $\mathbb{B}$ sur un domaine $\mathbb{D}'$ (non nécessairement entièrement distinct de $\mathbb{D}$) et avec les contraintes exprimées par la théorie $\mathbb{C}$ sur le domaine $\mathbb{D}''$ et ainsi de suite.

Au niveau syntaxique, on pourrait, par exemple, disposer

(a) d'une représentation relativement adéquate des phénomènes $\mathbb{DV}$ 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* $\mathbb{GLF}$;

(b) d'une représentation des phénomènes de CLITICISATION PRONOMINALE $\mathbb{CP}$ dans le cadre de la théorie *gouvernement-liage* $\mathbb{GL}$;

et

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

Dès lors, on pourrait vouloir intégrer ces trois fragments dans un système $\mathbb{S}$:

(17)

$$\mathbb{S} = s(\mathbb{GLF}(\mathbb{DV})) \;\&\; s(\mathbb{GL}(\mathbb{CP})) \;\&\; s(\mathbb{GSG}(\mathbb{COC}))^{24}$$

Comment intégrer ces différentes descriptions, les rendre compatibles l'une avec l'autre et les exploiter dans un système efficace et transparent?

Ce que nous nous proposons de construire, c'est un cadre métathéorique 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.

La définition de ces domaines et l'élaboration de (méta)formalismes d'expression de *contraintes* (les GRAMMAIRES DE CONTRAINTES) constitueront un autre objectif fondamental de recherche dans les années à venir.

24. s est ici une fonction générique d'interprétation de l'ensemble $\mathbb{T}(\mathbb{D})$ des énoncés de la théorie $\mathbb{T}$ sur le domaine $\mathbb{D}$ dans $\mathbb{S}$.

Problèmes d'intégration inter-niveaux : interfaces

Niveaux

C'est un fait bien connu que les connaissances linguistiques ne sont pas monolithiques, mais organisées en plusieurs articulations, modules ou niveaux. Chacun de ces niveaux porte sur un ensemble d'objets et de relations spécifiques, ce qui implique des représentations de types essentiellement différents. Or, même dans une conception en *relais* (comme celle de Jackendoff (1987)), où chaque niveau n'est interfacé qu'aux niveaux immédiatement voisins, plutôt qu'à l'ensemble des autres niveaux, l'intégration des niveaux pose autant de problèmes que la représentation du contenu de chacun d'entre eux.

En d'autres termes, le contenu des « flèches » dans un schéma comme le suivant est tout aussi complexe que celui des objets qu'elles relient.

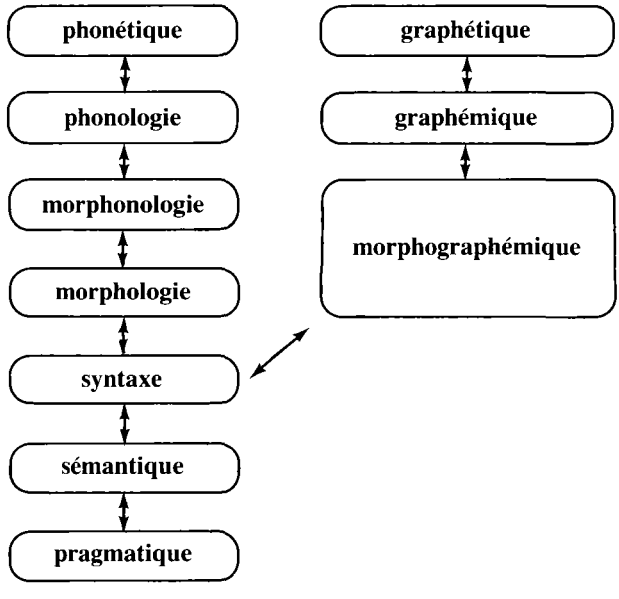


Figure 1 : Interaction des composantes dans une structure séquentielle

Par ailleurs, il semble bien que le lexique doive être interfacé à plusieurs niveaux différents, sinon à tous les niveaux de représentation linguistique.

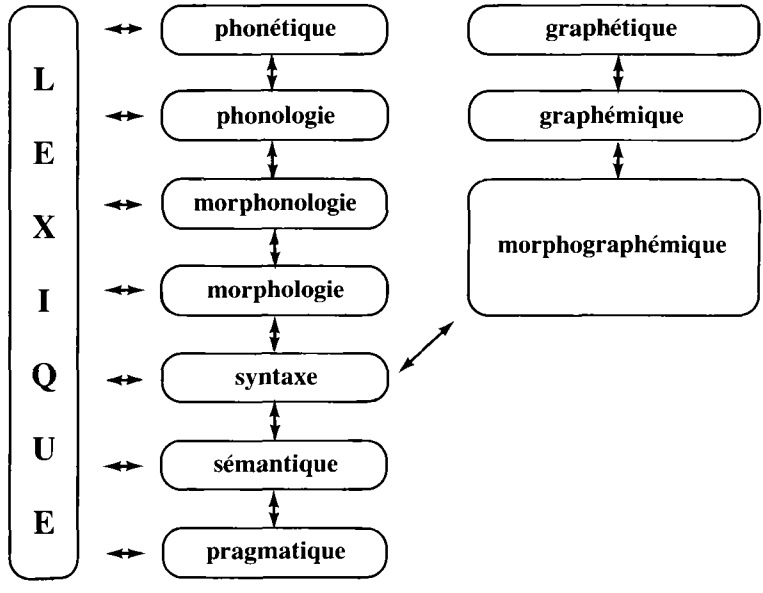


Figure 2 : Intégration du lexique aux autres composantes dans une structure séquentielle

Pis encore, dans un système hétérarchique, où chacun des niveaux peut interagir avec chacun des autres, comme les systèmes à dazibao, les problèmes d'interfaçage sont décuplés.

Comme le flot d'informations est beaucoup plus important entre modules voisins, on adopte généralement un modèle séquentiel à relais, plutôt qu'un modèle à dazibao, quitte à exprimer les autres contraintes au moyen d'extensions *ad hoc*.

Quoi qu'il en soit, l'immense majorité des systèmes expérimentaux existants ne porte que sur un ou deux de ces niveaux. Il y a donc là un gigantesque travail préliminaire à accomplir.

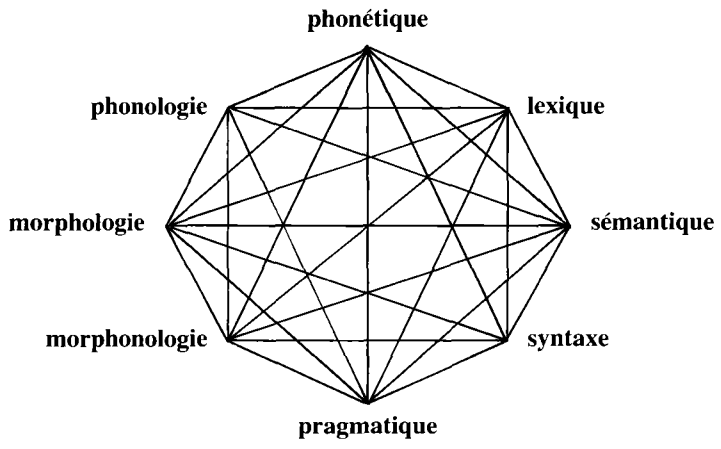


Figure 3 : Intégration des composantes dans une structure hétérarchique

Objets et interfaces

C'est ici que la méthodologie des représentations-objets est particulièrement intéressante. On peut concevoir chaque module comme une classe d'objets. Les classes sont elles-mêmes des objets qui communiquent avec les autres objets au moyen de messages. La façon dont un message est interprété par son récepteur dépend uniquement du contenu du message et des propriétés intrinsèques du récepteur.

On peut donc, en principe, définir à l'intérieur même d'un module à quels types de messages les objets du module peuvent réagir et comment ils peuvent y réagir.

Nous en sommes actuellement à concevoir des maquettes réalisant ce type d'interfaçage.

Intégration inter-représentations

Représentation des connaissances linguistiques et représentations de représentations

Métalangage de représentation

Toute théorie linguistique définit, implicitement ou explicitement, un métalangage de représentation des connaissances linguistiques doté d'une syntaxe et d'une sémantique propres.

Syntaxe

La syntaxe d'un métalangage de représentation des connaissances linguistiques est l'ensemble des expressions bien formées de ce métalangage.

Si, comme c'est le cas pour la théorie gouvernement-liage par exemple, le métalangage n'est pas explicitement défini, mais est construit à partir d'une langue naturelle quelconque, il est nécessaire de reconstruire ce métalangage de façon formelle.

Il s'agit tout d'abord d'établir le vocabulaire, c'est-à-dire les prédicats théoriques explicites de la théorie (par exemple *C-COMMANDE* , *GOUVERNE*, *ASSIGNE UN CAS*, *ASSIGNE UN θ-RÔLE*, etc.) et les prédicats théoriques implicites sous-jacents à la définition de ceux-ci (par exemple, *domine*, *est étiqueté*, *précède*, etc.) et de construire une syntaxe explicite et rigoureuse pour ceux-ci.

À titre indicatif, voici une typologie naïve et non exhaustive, des prédicats explicites communs à différentes variantes de la théorie gouvernement-liage et des modalités pouvant les affecter.

Dans une théorie comme la grammaire syntagmatique généralisée, qui est basée sur un ensemble de prédicats théoriques explicites (prédicats élémentaires : *SUBSUME*, *UNIFIE*, *a pour PORTÉE*, *a pour VALEUR*, *est une EXTENSION*, *une ABSTRACTION*, *une FUSION*, *une GÉNÉRALISATION*, prédicats-règles : *DI*, *PL*, *MÉTA*, *RCT*, *STD*, et prédicats principes universels : *PTT*, *PTP*, *PAC*), cet inventaire des prédicats théoriques sous-jacents est évidemment simplifié[25], comme est simplifiée la détermination de leur signature.

À ces prédicats, il faut ajouter les connecteurs logiques : *non* ($\neg$), *et* ($\wedge$), *ou* ($\vee$), *implique* ($\supset$) et *si et seulement si* (ssi, <=>), munis de leur interprétation habituelle. Dans cette liste de prédicats, la première forme entre parenthèses est une notation préfixée de la relation, la deuxième est la notation généralement employée[26].

---

25. *Cf.* Morin (1989) et (1989a) sur les prédicats théoriques de GSG et leur exportation dans des domaines d'interprétation externes.
26. Les symboles employés ont la signification suivante :

(i) α, α', β, β' (avec ou sans indice) sont des **catégories**, c'est-à-dire des ensembles de **spécifications de traits** $\{[t_i : v_i], [t_j : v_j], [t_k : v_k] \ldots\}$;

(ii) **t** est un **trait**, $\mathbf{t} \in \mathbf{T}$ (un ensemble fini de symboles, les **traits**);

(iii) **v** est une **valeur**, $\mathbf{v} \in \mathbf{V}$ (un ensemble fini de **valeurs**, atomiques ou complexes),
$\mathbf{V} = \mathbf{V}_{\mathbf{ATOMIQUES}} \cup \mathbf{V}_{\mathbf{COMPLEXES}}$
$\mathbf{V}_{\mathbf{ATOMIQUES}}$ = un ensemble fini de symboles, disjoint de **T**, les **valeurs atomiques**
$\mathbf{V}_{\mathbf{COMPLEXES}}$ = l'ensemble **C** des catégories
$\mathbf{V}'$ est un sous-ensemble de **V** ($\mathbf{V}' \in \wp(\mathbf{V})$);

(iv) **W** est une variable pour un multi-ensemble quelconque de catégories (c'est-à-dire un ensemble où un même élément peut apparaître plusieurs fois. Par exemple, A_m = {a, b, a, c}$_m$ est un multi-ensemble construit sur l'ensemble de base A = {a, b, c});

(v) $\wp(\mathbf{V})$ est l'ensemble des sous-ensembles de **V**.

PRÉDICATS

C-Commande
c-commande, appartient au domaine de
c-commande de

ACCESSIBILITÉ
est accessible, adjacent à

GOUVERNEMENT
gouverne,
est une catégorie gouvernante/de gouvernement (pour)

CONTRÔLE
contrôle

LIAGE
A/~A–lie, est A/~A–lié, est A/~A–libre,
est en relation de liage mutuel avec,
est une catégorie de liage,

est { *coindicé* / *coréférentiel* / *cosuperindicé* } *à*

est { *une anaphore* / *un pronominal* / *une variable* / *une r-expression* }

CONSTITUANCE, DOMINATION, PROJECTION
domine, contient, est un constituant immédiat de,
est une projection

SÉLECTION
sélectionne

POSITIONS
est / occupe une A/~A, θ/~θ, K/~K–position,
est une position génitive / prénominale

NIVEAUX
est lexical, est un X^O, X^I

BORNAGE, BARRIÈRES
est une borne / barrière

CAS, Θ-RÔLES
assigne un Cas / θ-rôle à, K-/θ-marque,
reçoit un Cas/θ-rôle de,
est θ-marqué par, absorbe le Cas (C) (de),
retire la possibilité d'assigner le Cas à

CHAÎNES
est une chaîne / un chemin ambigu/non ambigu, brise une chaîne / un chemin

VISIBILITÉ, MORPHOLOGIE, PHONÉTIQUE
est visible / doté d'une matrice phonétique, possède une matrice de traits intègre,
est un AGR « riche », etc.

EFFACEMENT DANS COMP
efface ($[_{N^-}$ *e]*), *assigne un indice à (Comp).*

MODALITÉS

NIVEAUX
en D-/S-structure, en FL, en FP,
en structure de surface,
à tous les niveaux de dérivation, au niveau N_i

GOUVERNEMENT
par gouvernement

DOMAINE LINGUISTIQUE
dans la (famille de) langue(s) L_j

LOCALITÉ
localement

MODALITÉS DIVERSES
lexicalement
structuralement
thématiquement
facultativement
proprement
minimalement/maximalement
directement/indirectement

Figure 4 : Prédicats et modalités de la théorie GL

Prédicats élémentaires

a) *a pour portée*	(PORTÉE [t, V']	ou	ρ (t) = V'	),
b) *a pour valeur*	(VAL [t, v].	ou	[t : v]	),
c) *subsume*	(SUBS [α, α'],	ou	$\alpha \subseteq \alpha'$	),
d) *est une extension de*	(EXT [α, α'],	ou	$\alpha \supseteq \alpha'$	),
e) *unifie*	(UNI [α, {β_1, β_2, ..., β_n}]	ou	(α, {β_1, β_2, ..., β_n})	),
f) *généralise*	(GÉN [α, {β_1, β_2, ..., β_n}]	ou	(α, {β_1, β_2, ..., β_n})	),
g) *est une abstraction de*	(ABS [α, {β_1, β_2, ..., β_n}]	ou	Λ (α, {β_1, β_2, ..., β_n})	),
h) *est une fusion de*	(FUS [α, {β_1, β_2, ..., β_n}]	ou	Φ (α, {β_1, β_2, ..., β_n})	).

Figure 5 : Prédicats élémentaires de GSG

Prédicats-règles

a) *domine immédiatement*	(DI [α, {β_1, ..., β_n}]	ou	$\alpha \rightarrow \beta_1, \ldots, \beta_n$	),
b) *précède linéairement*	(PL [α, β]	ou	$\alpha \ll \beta$	),
c) *RCT*	(RCT [α, β]	ou	RCT : $\alpha \supset \beta$	),
d) *STD*	(STD [α, β]	ou	STD : $\alpha \supset \beta$	),
e) *métarègle*	(MÉTA $\left[\begin{array}{l} <\alpha, <\beta, W \gg, \\ <\alpha', <\beta', W \gg \end{array}\right]$	ou	$\begin{array}{ll} \alpha & \beta, W \\ \alpha' & \beta', W \end{array}$	).

Figure 6 : Prédicats-règles de GSG

Principes universels d'instanciation

a) *Principe de traits de tête*
(PTT [α, {β_1, ..., β_n}, {β_1, ..., β_k}] ou PTT [α, (β_1, ..., β_1, ..., β_k, ..., β_n)]).

b) *Principe de traits de pied* (PTP [α, {β_1, ..., β_n}] ou PTP [α (β_1, ..., β_n)]),

c) *Principe d'accord-contrôle* (PAC [α {β_1, ..., β_n}] ou PAC [α (β_1, ..., β_n)]).

Figure 7 : Prédicats-principes universels de GSG

Voici les signatures de ces différents prédicats (c'est-à-dire les types des arguments qu'ils prennent et, dans le cas des fonctions, les types des valeurs qu'elles retournent)[27].

Prédicats-élémentaires

PORTÉE :	$T \longrightarrow \wp(V)$	(fonction)
VALEUR :	$T \longrightarrow V$	(fonction)
SUBSUME :	$((T \longrightarrow V), (T \longrightarrow V))$	(relation)
EXTENSION :	$((T \longrightarrow V), (T \longrightarrow V))$	(relation)
UNIFIE :	$\wp(T \longrightarrow V) \longrightarrow (T \longrightarrow V)$	(fonction)
GÉNÉRALISATION :	$\wp(T \longrightarrow V) \longrightarrow (T \longrightarrow V)$	(fonction)
ABSTRACTION :	$\wp(T \longrightarrow V) \longrightarrow (T \longrightarrow V)$	(fonction)
FUSION :	$\wp(T \longrightarrow V) \longrightarrow (T \longrightarrow V)$	(fonction)

Prédicats-règles

DI :	$((T \longrightarrow V), \wp(T \longrightarrow V))$	(relation)
PL :	$((T \longrightarrow V), (T \longrightarrow V))$	(relation)
RCT :	$((T \longrightarrow V), (T \longrightarrow V))$	(relation)
STD :	$((T \longrightarrow V), (T \longrightarrow V))$	(relation)

Principes universels d'instanciation

$((T \longrightarrow V), \wp(T \longrightarrow V)) \longrightarrow ((T \longrightarrow V), \wp(T \longrightarrow V))$ (fonction)

Figure 8 : Signatures des prédicats théoriques de GSG

Sémantique

Mais un tel travail ne suffit pas, encore faut-il donner une sémantique descriptive rigoureuse à chacun de ces prédicats. Il faut définir les domaines d'objets caractérisés et les relations que les prédicats théoriques y définissent. En d'autres termes, il s'agit de typer les prédicats relativement aux objets des domaines qu'ils servent à définir.

Ce travail, loin d'être trivial, est absolument nécessaire, bien qu'il soit la plupart du temps négligé. En effet, les « modèles informatiques » des théories linguistiques ne partagent le plus souvent avec ces dernières qu'un sous-ensemble non interprété du vocabulaire théorique.

27. On aura noté que la PORTÉE, la VALEUR, l'UNIFICATION, la GÉNÉRALISATION, l'ABSTRACTION et la FUSION ne sont pas des relations, mais des opérations (fonctions), que nous transformons en relations en ajoutant comme argument leur résultat. En fait, chacun des prédicats théoriques peut être reconstruit comme une fonction caractéristique retournant l'ensemble des objets pour lesquels il est vrai dans une interprétation donnée. Pour plus de simplicité, nous n'avons pas donné cette reconstruction ici.

Un tel travail d'identification de la *signature* des théories linguistiques est actuellement en cours pour la théorie gouvernement-liage et pour la théorie syntagmatique généralisée.

Domaines et dimensions

L'intégration des représentations que nous proposons est une intégration sémantique plutôt que syntaxique. C'est-à-dire que c'est à partir des domaines définis par les différentes théories (métasémantique), plutôt qu'à partir de leur structure interne (métasyntaxe) que nous tentons de les intégrer. Pour ce faire, il est nécessaire de procéder à un repérage *a priori* des dimensions de l'espace définissable pour chacun des niveaux de façon à pouvoir vérifier comment chaque théorie particulière recouvre cet espace, comment elle y établit son domaine. On peut, par exemple, relativement indépendamment de toute théorie syntaxique ou sémantique, définir les dimensions des espaces syntaxique et sémantique pour les langues naturelles.

Dimensions syntaxiques

Le domaine syntaxique peut être défini par les dimensions suivantes.

Structure syntagmatique

CATÉGORIES

ORDONNANCEMENT

HIÉRARCHIE

Dépendances

GOUVERNEMENT	LIAGES
sélection	anaphore
rection	coréférence
	quantification
	accord
FILTRAGES	LIAGES-FILTRAGES
répulsion	contrôle
absorption	« mouvements »
haplologie	(*arguments, opérateurs, parasites*)
anaccord	
exclusions	

BORNAGES

Globaux/locaux Horizontaux/verticaux

Figure 9 : Dimensions syntaxiques

Dimensions sémantiques

Le domaine sémantique peut se définir par les dimensions suivantes.

θ-structure
- Rôles des arguments
  - AGENT
  - PATIENT
  - LOCATION
  - SOURCE
  - CIBLE
  - ...

ρ-structure
- Rôles discursifs
  - THÈME-RHÈME
  - PERTINENCE / CENTRALITÉ
  - CENTRE / PÉRIPHÉRIE
  - NOUVEAU / CONNU
  - EMPHATIQUE / RETIRÉ,
  - FOCUS / PRÉSUPPOSITION

μ-structure
- Modalités
  - NÉCESSITÉ / POSSIBILITÉ
  - OBLIGATION / PERMISSION
  - PROBABILITÉ / CERTITUDE

ι-structure
- Structure intensionnelle
  - QUANTIFICATION
  - DÉTERMINATION
  - MODULATION
  - LIMITATION

τ-structure
- Temps et aspect

ε-structure
- Espace et lieux
- Parties et limites

π-structure
- Structure locutoire et performative

Figure 10 : Dimensions sémantiques

Les limites d'espace nous empêchent d'expliciter ici cet inventaire de dimensions syntaxiques et sémantiques et de montrer comment chaque théorie syntaxique ou sémantique doit pouvoir y être interprétée.

Conclusion : institutions grammaticales

Au fond, ce que nous devrons essayer de caractériser, avant de pouvoir interfacer des REPRÉSENTATIONS, ce sont des INSTITUTIONS GRAMMATICALES, c'est-à-dire des espaces, des cadres généraux indépendants des représentations liées à des théories linguistiques particulières. Par ailleurs, ces INSTITUTIONS GRAMMATICALES[28], définies selon leurs dimensions, permettront également de mieux voir la

28. Sur la notion formelle d'*institution* en logique des programmes, voir Goguen & Burstall (1983, 1985, 1986).

nature de chacun des NIVEAUX de représentation et comment ils peuvent être interfacés à leur tour les uns aux autres.

Nous sommes encore bien loin de cet idéal, mais les développements récents sont particulièrement encourageants. Le génie linguistique (comme, avant lui, le génie logiciel) ne sortira du bricolage qu'au prix d'une réflexion approfondie et rigoureuse sur la sémantique des modèles qu'il utilise.

Références

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, p. 114-135.

BARRETT E. (1988) : *Text, ConText and HyperText : Writing with and for the Computer*, Cambridge, MA, MIT Press.

BARRETT E. (1989 réd.) : *The Society of Text : Hypertext, Hypermedia, and the Social Construction of Information*, Cambridge, MA, MIT Press.

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., L. EMIRKANIAN 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.

DA SYLVA L. (1990) : *Un parseur inspiré de la théorie gouvernement-liage*, mémoire de M.A. inédit, département de linguistique, philologie et traduction, Université de Montréal.

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

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

GAZDAR G., G. PULLUM, R. CARPENTER, E. KLEIN, T. HUKARI & 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, p. 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 M. (1985) : *Méthodes en syntaxe*, Paris, Hermann.

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

MEL'ČUK I. *et al.* (1984, 1988) : *Dictionnaire explicatif et combinatoire du français, Recherches lexico-sémantiques I et II*, Montréal, Presses de l'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. (1989) : *Syntaxe*, département de linguistique et philologie, Université de Montréal.

MORIN J.-Y. (1989a) : « Prédicats théoriques et données externes. Syntaxe diachronique », *Revue canadienne de linguistique*, numéro spécial, *Linguistic Theory and External Evidence.*, 33, 4 : 443-75.

MORIN J.-Y. (1989b) : « Particules et parsage universel », in Weydt, H. (1989 réd.) *Sprechen mit Partikeln*, Berlin, De Gruyter, p. 713-728.

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

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* : 87-94.

SAINT-DIZIER P. et 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, n° 32.

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

11

Poste de travail de traduction par l'informatisation d'un dictionnaire français-arabe

Saadia TALEB

Institut d'Études et de Recherches pour l'Arabisation (IERA), Université Mohamed V, Rabat, Maroc

La réalisation d'un poste de travail de traduction par l'informatisation d'un dictionnaire français-arabe compatible avec un traitement de texte arabisé grand public se situe dans le cadre d'un contrat entre l'ACCT et l'association Champollion. Ce contrat associe l'IERA en tant qu'intervenant apportant une compétence spécifique en matière de mise au point de bases de données lexicographiques et terminologiques multilingues.

Combinant différentes aides informatiques, le poste de travail pour le traducteur de français vers l'arabe a pour objectif d'automatiser certaines tâches connexes à l'acte traductionnel.

Ce poste est installé sur petit matériel, muni d'affichage graphique (Macintosh) et largement diffusé, et repose sur le principe de l'intégration d'un ensemble d'outils informatiques et de savoir-faire réellement existant :

- Un traitement de texte Wintext qui offre la possibilité d'associer plusieurs langues dans le même document et permet de synchroniser deux fenêtres dans lesquelles apparaîtront le texte source et sa traduction. Le traitement de texte dispose d'un correcteur orthographique français qui utilise un dictionnaire de 246 000 mots avec la possibilité de mettre au point son propre lexique ainsi que celle d'exploiter les données de documents externes crées par d'autres applications, notamment Winfile et WinTool.

- Un lemmatiseur (ODILE) pour trouver automatiquement la forme canonique d'un mot en évitant de forcer l'utilisateur à taper cette forme. Il permet de fabriquer, à partir d'une chaîne sélectionnée, un ensemble de clés de recherche, de les compléter si nécessaire « à la main » et de s'en servir pour accéder aux dictionnaires gérés par WinTool.
- L'accès à un dictionnaire informatisé.

La recherche des équivalents est une opération essentielle de l'acte traductionnel. L'informatisation de cette tâche augmente la productivité du traducteur. Cette productivité augmente d'autant plus que le dictionnaire est modifiable et peut être mis à jour constamment par l'utilisateur.

L'outil dictionnaire du poste de travail du traducteur est pris en charge par l'utilitaire WinTool qui est un accessoire de bureau disponible dans le menu « pamno » du Macintosh à partir de l'application Wintext. Cet utilitaire permet d'ouvrir deux bases simultanément.

L'outil dictionnaire mis au point dans le cadre de ce travail est un dictionnaire français-arabe de 15 000 termes spécialisés dans les domaines des finances : comptabilité et banque. Ces termes sont extraits de la base de données LEXAR.

Cette base de données lexicographiques et terminologiques réalisés à l'IERA sur un IBM 4341, constitue un dictionnaire multilingue géant et se présente comme un instrument de référence lexicographique et terminologique de première importance et par ses dimensions (quelque 700 000 fiches) et par la diversité des sources à partir desquelles a été puisée la matière linguistique. Cette matière à la fois complexe et hétérogène subit une transformation lors d'une opération d'analyse qui vise à représenter le contenu du document primaire sous une forme différente en respectant un certain nombre de règles prédéfinies. Ces règles représentent un nombre précis pour permettre la distribution de la matière linguistique selon des catégories prédéterminées aisément reconnaissables par la machine et assez générales pour que, quelle que soit la complexité du dictionnaire ou du lexique à stocker, ce dernier puisse s'intégrer à l'ensemble et être réduit aux mêmes catégories prévues.

La donnée LEXAR est de ce point de vue une entité indépendante ayant ces propres spécifications et pouvant être réutilisée pour différentes applications.

Pour réaliser cet outil dictionnaire sur l'utilitaire WinTool, il a fallu sélectionner les données LEXAR relatives aux domaines des spécialités choisies, les extraire de l'IBM central, les transférer sur micro-ordinateur PS/2, procéder à un décodage et à un réencodage selon le codage arabe utilisé sur Macintosh, et enfin, définir une base de données WinTool et y transférer les données.

Cette opération a montré la possibilité de tirer de la base de données LEXAR des dictionnaires spécialisés de 7 000 à 8 000 termes et leurs équivalents pouvant tenir sur une disquette de 800 Ko, ou de créer une collection de tels dictionnaires sur CD-ROM.

On peut donc personnaliser le poste de traduction en offrant des dictionnaires spécialisés selon les besoins de tel ou tel utilisateur.

Il s'agit là d'une première expérimentation et il va sans dire que les outils d'aides doivent être encore plus perfectionnés et que l'on peut intégrer d'autres aides infor-

matiques, notamment un vérificateur orthographique pour l'arabe, un système de conjugaison et un logiciel de fenêtrage performant capable de gérer efficacement et simultanément toutes ces applications. Ces logiciels doivent répondre à une double exigence : une convivialité extrême et une haute performance.

12

Traduction et terminologie : expérience et perspectives industrielles

Daniel BACHUT, René GERBER[1]

SITE, Maisons-Alfort; B'VITAL, Grenoble, France

Résumé

Le groupe SITE, leader en ingénierie documentaire (documentation, traduction, logistique, etc.), développe aujourd'hui plusieurs activités, qui se situent au cœur même des domaines de la lexicographie, de la terminologie et de la traduction. Dans cet article, nous décrivons ces différentes activités, puis nous présentons les conditions techniques nécessaires à la conception du poste de travail du rédacteur-traducteur de l'an 2001.

La traduction

Le département de traduction

Le département de traduction de SITE compte aujourd'hui 60 traducteurs, qui traitent plus de 110 000 pages par an. C'est un département à croissance rapide, conséquence des besoins de plus en plus pressants du marché sur ce créneau. Les langues les plus couramment traitées sont le français et l'anglais, et dans une moindre mesure l'allemand

1. Les auteurs tiennent à associer à la rédaction de cet article Mme B. Roudaud, M. J.-J. Perot et M. D. Maret, sans l'aide desquels il aurait été difficile de réaliser un tour d'horizon aussi complet des activités de la société.

et l'espagnol. Il s'agit essentiellement de traduction technique dans des domaines tels que l'aéronautique, l'informatique, la mécanique, etc.

Les traducteurs travaillent dans un environnement informatisé moderne. L'organisation informatique de base consiste en un réseau de PC/PS, connectés par Token Ring. Différents traitements de texte sont disponibles, le plus couramment employé étant le logiciel Word de Microsoft. Le poste de travail donne accès à la base terminologique multilingue PHENIX, présentée plus loin. La variété des besoins des clients amène également les traducteurs à travailler dans d'autres types d'environnement : Macintosh et, depuis peu, stations de travail sous UNIX.

L'objectif de qualité conduit les équipes à pratiquer souvent la lecture croisée en guise de révision. Plusieurs types de relecture peuvent être effectués : relecture linguistique (orthographique et stylistique), relecture technique (par un traducteur spécialisé ou un spécialiste du domaine). L'aspect terminologique est pris en compte depuis de nombreuses années, dans un souci de productivité, d'homogénéité, de qualité, et même de formation. À cet égard, on estime que, depuis l'utilisation de la base terminologique PHENIX, le temps de formation d'un traducteur à une langue de spécialité est passé de deux ans à six mois. Souvent le client impose l'utilisation d'une *terminologie approuvée*, définie par lui, que les traducteurs doivent s'engager à respecter.

Ajoutons que les demandes des clients peuvent être très variées. Certains ont des exigences telles, que la traduction consiste quasiment en la réécriture du document en langue cible (traduction de l'anglais simplifié en français « normal », par exemple). D'autres demandent, au contraire, une traduction très fidèle du point du vue stylistique, c'est-à-dire aussi proche que possible du texte source. Cela est justifié pour des documents appelés à être modifiés, et pour lesquels le client tient à conserver un certain parallélisme entre les textes source et traduit pour simplifier les mises à jour.

La traduction assistée par ordinateur

L'industrialisation du système ARIANE F → A

Début 1990, SITE a pris une part majoritaire dans le capital de la société grenobloise B'VITAL, spécialisée dans le développement de systèmes de traduction assistée par ordinateur (TAO), et dont l'une des réalisations majeures consiste dans le développement d'un système de traduction français-anglais de deuxième génération, appelé ARIANE F → A. Ce système utilise le générateur de systèmes de TAO ARIANE, mis au point par le laboratoire GETA (Groupe d'études pour la traduction automatique). Les linguiciels (grammaires et dictionnaires) d'ARIANE F → A ont été conçus pour traiter des textes techniques, en particulier dans le domaine aéronautique.

Un projet d'industrialisation d'ARIANE F → A est actuellement mené par les trois partenaires (SITE, B'VITAL et le GETA), avec le soutien du ministère de l'Industrie. Il a pour objectif de faire passer l'outil d'un état de prototype avancé à un

état industriel. Pour y parvenir, les partenaires ont préconisé différentes actions, que l'on peut regrouper suivant deux axes principaux :

- l'« habillage » du système, par la réalisation des outils connexes permettant de l'intégrer dans une chaîne documentaire industrielle;
- la mise en service du système dans des conditions réelles, en lui faisant traduire plusieurs milliers de pages. Évaluation du système en ce qui a trait à la qualité des traductions, au coût, au délai et au volume.

Habillage du système

ARIANE fonctionne sur *mainframe* IBM, et propose aux linguistes un environnement de développement très riche. En exploitation, l'utilisateur final voudra disposer de son traitement de texte habituel (sous MS-DOS par exemple) et bénéficier d'un grand confort au niveau de l'interface (menus déroulant, multifenêtrage, souris, etc.). Un premier moniteur d'exploitation a donc été réalisé sous MS-DOS, pour faciliter l'accès au système de TAO, transformé alors en serveur de traduction. Cet environnement permet à un utilisateur de lancer sa traduction de manière transparente, sans qu'il lui soit nécessaire de savoir où celle-ci a réellement lieu. De plus, la traduction étant déportée, l'utilisateur peut continuer à travailler sur son micro-ordinateur. Actuellement la traduction peut être lancée à partir de MS-DOS, ou directement sous Word par l'intermédiaire d'un jeu de touches de fonctions.

Un travail important a également été réalisé pour permettre la restitution du format du texte source dans le texte traduit, dans le but de diminuer sensiblement la charge de travail du post-éditeur. Une représentation « normalisée » des ordres de formatage et de structuration des textes a été définie et prise en compte par les processus linguistiques. Des pré et post-processeurs, complètement intégrés au nouveau moniteur, ont été réalisés pour passer de la forme des textes d'entrée à la forme « normalisée », et de cette forme à la forme de sortie. Une généralisation de cette représentation en SGML est prévue.

Évaluation du système et premiers résultats

Les premiers tests sur la documentation de l'avion FALCON 900 ont été réalisés sans la terminologie spécifique, pour des raisons de délai et de disponibilité. La présence dans les dictionnaires d'ARIANE F $\longrightarrow$ A de nombreux termes du domaine aéronautique (8000 environ) a cependant rendu possibles ces essais. Notons que la récupération automatique du glossaire spécifique au FALCON 900 (articles PHENIX) vient d'être achevée.

Les traducteurs ont considéré que les traductions obtenues étaient de bonne qualité. Les calculs réalisés sur les temps de post-édition donnent aujourd'hui une demi-heure par page révisée pour les cas les plus défavorables, permettant d'améliorer la productivité des traducteurs par un facteur deux (16 pages par jour et par traducteur au lieu de 8 pages classiquement). Il est certain que l'ajout de la terminologie permettra de réduire encore le temps de révision.

L'amélioration de la productivité (par diminution des délais) ne se traduit cependant pas par une diminution comparable des coûts, puisqu'à celui du réviseur, il faut ajouter le coût des ressources informatiques nécessaires, ainsi que le coût du système lui-même. Le coût de l'amortissement du système est faible, dans la mesure où un système de TAO de ce type ne peut intéresser qu'un client ayant des besoins importants (plusieurs dizaines de milliers de pages par an).

Ce sont les ressources informatiques (essentiellement du CPU) qui représentent aujourd'hui l'élément de coût le plus important pour une traduction réalisée avec ARIANE F ⟶ A. Cela tient au fait que l'implémentation est dépendante de l'architecture IBM 370. Remarquons que le coût de la seconde CPU peut varier du simple au quadruple suivant les centres de calcul (toutes choses étant égales par ailleurs). Ces premières conclusions (bonne qualité, productivité double mais coûts informatiques élevés) ont motivé SITE à préparer le projet EUROLANG.

Le projet EUROLANG

En collaboration avec plusieurs partenaires industriels et universitaires européens, SITE a déposé un projet EUREKA fin mars 1991. Le projet EUROLANG a pour objectif la réalisation d'un système de TAO multilingue fonctionnant sur cinq langues européennes (français, anglais, allemand, espagnol et italien). Pour construire ce système, on développera une boîte à outils pour le traitement automatique de la langue naturelle (TALN).

Le principe d'une boîte à outils donne la possibilité d'utiliser une stratégie *meccano* (*plug and play*), qui offre des garanties d'ouverture tant au niveau informatique (adéquation de chaque outil, possibilité d'assemblage) qu'au niveau linguistique (mise en jeu de plusieurs théories, stratégies définies par les utilisateurs de la boîte à outils).

Le projet est bâti autour d'un noyau de partenaires représentant les différents centres de compétences européens en matière de TAO : SYSTRAN, SITE, le laboratoire GETA et plusieurs équipes de recherche participant au projet EUROTRA. La participation de sociétés déjà utilisatrices de la TAO, comme RANK XEROX (UK) et KRUPP (RFA), donne au projet une dimension industrielle. Les différents partenaires, qui travaillaient jusqu'alors sur des projets concurrents, sont parvenus à la conclusion que la mise en commun des hommes, des techniques, des méthodes et des idées était le seul moyen d'aboutir à un système de TAO à la pointe de la technologie, et capable de faire face aux énormes efforts consentis par les Japonais dans ce domaine.

La terminologie et la lexicographie

En terminologie et lexicographie, SITE mène différentes activités concernant aussi bien la production de glossaires terminologiques et de lexiques, que la conception et la réalisation d'outils d'enrichissement, de consultation et de stockage de dictionnaires électroniques.

PHENIX, la base terminologique multilingue à l'usage des traducteurs

La base PHENIX, conçue et réalisée en interne, permet de gérer des fiches terminologiques constituées, pour chaque mot (ou expression), de sa(ses) traduction(s), de son contexte d'utilisation (domaine, sous-domaine, etc.), de quelques informations linguistiques (catégorie, genre, etc.), et d'un *code objet*. Ce dernier est en quelque sorte une caractérisation sémantique, véhiculant des informations précieuses pour un traducteur. On distingue environ 200 codes objets dans PHENIX (par exemple, *anomalie, métal, pays*).

Un système de gestion évolué permet aux traducteurs d'accéder aisément à cette base. On peut notamment exprimer des critères de sélection très raffinés pour trouver l'équivalent approprié dans un contexte donné (domaine, client, composant, etc.). L'accès est possible à partir de n'importe quelle langue (aucune langue n'est privilégiée comme source). Aujourd'hui, la base PHENIX comporte environ 240 000 entrées multilingues (français, anglais, allemand, espagnol) dans divers domaines techniques, en particulier l'aéronautique.

PHENIX fonctionne de plus en réseau, permettant ainsi le partage de l'expérience acquise à l'intérieur d'une équipe de traducteurs, qui l'enrichissent et la consultent. Pour assurer la qualité et l'homogénéité du contenu, deux terminologues spécialisés sont chargés de valider les nouveaux termes.

Remarquons que la terminologie, après avoir été introduite comme aide à la traduction, a acquis ses lettres de noblesse, puisqu'elle est vendue maintenant pour elle-même et non plus au titre d'une autre prestation. Cette prise de conscience de l'importance de la terminologie pour les entreprises ne se limite d'ailleurs pas à l'aspect traduction, mais englobe tous les aspects de la production de documents, et en particulier la rédaction.

BDTAO, la base lexicale dédiée à la TAO

Pour réduire les coûts des dictionnaires de TAO, tout en améliorant leur qualité, B'VITAL a mis au point une méthodologie d'enrichissement de ces dictionnaires. Le système BDTAO a été développé dans ce contexte pour permettre une gestion aisée et naturelle des données lexicales et la production des dictionnaires de TAO (système dédié à la TAO). Cette base sert à l'enrichissement et à la mise à jour des dictionnaires du système ARIANE F ⟶ A. BDTAO contient actuellement environ 20 000 termes français et anglais, à partir desquels sont produits automatiquement les différents dictionnaires (monolingues et bilingues) nécessaires au fonctionnement du système ARIANE. Les entrées de BDTAO sont très riches au niveau des informations morphologiques, syntaxiques et sémantiques.

Face au coût encore important de l'enrichissement manuel de BDTAO par les lexicographes, l'équipe a décidé de tenter une expérience de récupération automatique de données lexicales à partir de la base PHENIX : il s'agissait de charger automatiquement le glossaire du FALCON 900 dans BDTAO. Ce glossaire de 5000 entrées

bilingues, utiles à l'évaluation de ARIANNE F ⟶ A, a été chargé dans BDTAO avec un certain succès : environ 90 pour cent des termes ont été complètement et correctement convertis, notamment grâce à l'ajout d'un module de précalculs linguistiques dans le programme de conversion. Cette opération a permis d'économiser plusieurs mois-hommes par rapport à la méthode classique d'enrichissement manuel. La base PHENIX étant relativement riche (120 000 termes en français et en anglais), il devient donc possible de charger quasi automatiquement tous ces termes dans le système ARIANE F ⟶ A.

Un poste d'enrichissement lexicographique moderne

Le groupe SITE développe actuellement un poste d'enrichissement lexicographique et une base de données lexicales pour le CNET, dont les objectifs techniques sont proches de ceux affichés par les grands projets actuels du domaine, GENELEX et MULTILEX. Les développements sont en cours sur station de travail, et doivent aboutir à un système industriel fin 1991. Ce système est développé sous UNIX en respectant les standards actuels : C au niveau du langage de programmation, SQL au niveau de l'accès base de données, X-WINDOW/MOTIF pour ce qui est de l'interface homme-machine, SGML pour la définition du format d'échange des données, etc. À terme, un outil de ce genre sera utilisé comme base de référence pour le système de TAO.

Dans ce système, une entrée de la base lexicale est associée à des informations de natures très différentes : informations pour l'humain (définitions, exemples, etc.), informations pour le traitement automatique (transitivité, pronominalité, etc.). Ces informations sont modélisées sous forme d'attributs et de relations (relation d'abréviation, par exemple). Un réseau sémantique permettant de relier les concepts définis par les entrées lexicales est également mis à disposition du lexicographe.

Cette base est générique, dans la mesure où elle est indépendante d'une application ou d'une théorie linguistique particulière : l'utilisateur a les moyens de définir le type d'information qu'il désire associer à chaque terme, et de faire évoluer ces informations. Ainsi, la base n'est pas figée à un état donné de la technique. Cette caractéristique du système lui confère son indépendance vis-à-vis de la langue, et une évolution vers la prise en compte de données multilingues est envisagée.

Le poste de travail, ergonomique et productif, permet au lexicographe de saisir et de gérer toutes ses données. Ce poste met en œuvre toutes les techniques modernes de gestion et d'interfaçage (multi-fenêtrage, souris, boutons, icônes, etc.). Dans le but d'améliorer la productivité des lexicographes, le système met à disposition du lexicographe un L4G, lui permettant de définir :

- des accélérateurs linguistiques : il s'agit de précalculs linguistiques, fonction du lemme et de tous les champs déjà renseignés, qui permettent de préremplir un attribut ou de présélectionner un sous-ensemble de valeurs pertinentes dans une liste de valeurs possibles;

- des vérificateurs de cohérence : des fonctions permettant de vérifier différentes relations et dépendances inter-attributs seront activées soit en temps réel lors de l'enrichissement, soit en temps différé.

Cet outil générique permettra donc au CNET d'organiser, comme il l'entend, sa base de données lexicales. Il constitue en fait un point de départ très intéressant pour construire des dictionnaires électroniques répondant à différents besoins (terminologiques, lexicaux, dictionnaires du commerce, etc.). À partir de ces développements, SITE envisage de définir et de réaliser une version plus performante de PHENIX.

Vers un poste de travail intégré

Motivations et objectifs

Motivations

Les motivations pour un poste de travail moderne sont avant tout de type économique et technique. Il est clair que, avec la concurrence de plus en plus forte qui règne dans la plupart des grands domaines industriels (aéronautique, informatique, télécommunications, mécanique...), la qualité d'une documentation devient un facteur de décision important dans la prise de marché. Par exemple, la disponibilité de la documentation dans la langue du demandeur est un atout considérable.

Aujourd'hui, les coûts de constitution de la documentation sont souvent élevés, car celle-ci est souvent vue comme une activité annexe, mal intégrée au processus industriel, et toujours sous-évaluée (voire oubliée) lors de la définition d'un projet. Cet état des choses est heureusement en train de changer, et les industriels accordent de plus en plus d'importance à tous ces aspects : qualité de la rédaction et de la traduction, constitution d'une terminologie « normalisée », etc.

Des outils spécialisés puissants et adaptés

Dans le domaine de la production documentaire, il existe actuellement sur le marché un certain nombre d'outils spécialisés pour telle ou telle tâche. Citons, parmi les plus connus :

- les traitements de textes et les hypertextes,
- les outils dictionnaires : thésaurus, dictionnaires bilingues, dictionnaires de synonymes, etc.
- les outils de correction orthographique,
- les outils de gestion documentaire : indexation, interrogation (par mot clé ou autre),
- les outils de TAO : THAM (traduction humaine aidée par la machine) et TA (traduction automatique).

Notons que nombre de ces outils sont souvent difficiles à utiliser de manière simple dans les conditions réelles d'exploitation : par exemple, un rédacteur, travaillant sur un document technique français où il apparaît de nombreuses expressions anglaises (en informatique ou en aéronautique, par exemple), n'utilisera pas le correcteur orthographique français, qu'il jugera trop bavard et donc inefficace. De la même façon, un

outil de gestion documentaire ou un outil de TA, même de grande qualité, seront inutilisables s'il leur est impossible de gérer les ordres de mise en forme des documents.

Par ailleurs, de nouveaux outils commencent à apparaître (vérificateurs grammaticaux, vérificateurs stylistiques, analyseurs de corpus, etc.), qui étofferont la panoplie des utilisateurs. Les nouvelles possibilités qu'offre la technologie actuelle (processeurs rapides, réseaux distribués hétérogènes, bases réparties, écrans graphiques, CD-ROM, etc.) laissent penser que, dans 10 ans, le poste de travail aura subi des changements radicaux par rapport à son état actuel.

Une intégration poussée

La prolifération d'outils ne saurait être efficace que si l'ensemble de ces outils est intégré. Cela signifie que leur coexistence doit pouvoir être assurée sur un même poste de travail, et que la communication inter-outils est possible et transparente pour l'utilisateur final. En plus d'outils spécialisés tels que ceux présentés plus haut, le poste devra intégrer des outils plus classiques ou relevant d'autres techniques que celles qui nous intéressent ici : outils de DAO et de PAO, lecteur optique, serveur d'impression, lecteur de CD-ROM, accès à des bases de données textuelles et non textuelles (locales ou distantes), etc.

Il s'agit, en fait, de permettre à l'utilisateur d'effectuer toutes les opérations à partir du même environnement et de la façon la plus « économique » possible (sans refrappe...). Pour consulter un dictionnaire par exemple, il doit pouvoir sélectionner le terme dans le texte, puis activer la recherche. De même, un équivalent trouvé dans un dictionnaire doit pouvoir être inséré directement dans le fichier de travail (aussi simplement que s'il s'agissait d'un déplacement de texte).

C'est l'intégration de ces outils qui sera à la base des chaînes documentaires modernes, que les industriels visent pour la prochaine décennie. Cette démarche est analogue, et s'avère même complémentaire, à celle menée dans le cadre de l'initiative CALS du DoD, qui vise, dans un premier temps, à permettre la communication entre les diverses bases de données réalisées dans le cadre d'un projet industriel ou militaire (données de conception, de fabrication, de logistique, etc.), et, dans un second temps, à intégrer toutes ces données dans une base unique répartie (*publishing data base*), avec accès par vues multiples sur cette base.

Un poste multi-usage configurable

Le poste de travail (que nous avons jusqu'ici appelé *poste du rédacteur-traducteur*, pour simplifier nos propos) est, en fait, plus général que sa simple dénomination le laisse entendre (faut-il parler de *poste du communicateur*, comme certains le proposent?). Ce poste met en œuvre des outils qui sont communs à plusieurs métiers dans le domaine plus vaste de la gestion documentaire, et qui doivent procurer une aide pour différents types de tâches : rédaction, recherche terminologique, consultation et enrichissement de dictionnaires, traduction, pré et post-édition en TA, mise en page, etc.

Pour la définition du poste de travail, il faut éviter de tomber dans les pièges extrêmes qui sont :

- la définition d'un poste unique, utilisable par tous les métiers, mais qui sera trop riche et trop difficile à mettre en œuvre pour une utilisation particulière;
- la définition d'autant de postes qu'il y a de métiers, solution qui risque de conduire à une multiplication incontrôlable de configurations, étant donné l'évolution rapide des besoins et des métiers.

La définition d'un poste configurable qui offre à l'utilisateur toutes les fonctionnalités dont il a l'utilité, et celles-là seulement, apparaît comme la solution la plus ergonomique, et donc en définitive comme la plus productive. Le contrôle de l'activité sera laissé à l'utilisateur pour tout ce qui concerne les décisions de haut niveau : soumission de telle partie du texte à un système de TA, révision ou non du résultat, etc. Un tel poste permet de prendre en compte de manière souple les évolutions intervenant dans chaque métier, par adaptation de la configuration.

Mise en commun des compétences

Il est clair que la conception du poste de travail configurable nécessite non seulement l'intervention des divers métiers concernés (rédacteur, traducteur, terminologue, post-éditeur, etc.), mais également de spécialistes externes (ergonome, ingénieur en documentation, informaticien, spécialiste en bases de données, etc.). Remarquons qu'il ne s'agit ici que du développement du poste de travail, et non des divers outils qui y sont intégrés. Le coût de développement induit par cette entreprise est justifié par les enjeux que représente la mise au point d'une chaîne documentaire industrielle de bonne qualité.

Quelques fonctionnalités spécifiques au poste du traducteur

Pour terminer, nous décrivons un certain nombre de fonctionnalités, qui devraient être mises à la disposition du traducteur, utilisateur éventuel d'un système de TAO. Citons, par exemple :

- l'affichage simultané des textes source et traduit, avec des facilités de gestion synchronisée des deux fenêtres;
- les fonctions spécifiques : un traducteur, et en particulier un post-éditeur, appréciera certaines fonctions particulièrement adaptées à la manipulation textuelle (permutation intelligente de groupes de mots, rejet en fin de phrase, etc.). Ces fonctions auront une certaine connaissance de la structure textuelle;
- la manipulation grammaticales (mise au pluriel, changement du temps d'un paragraphe, mise à la voix passive/active, etc.);
- la recherche et le remplacement intelligents intégrant des connaissances grammaticales et typographiques;
- la détection des phrases répétitives, pour recherche automatique de leurs traductions;
- la gestion d'annotations successives (utiles pour les différentes relectures).

De manière générale, il s'agit d'introduire à tous les niveaux des fonctions intelligentes faisant appel à des processus linguistiques de différents types : lemmatiseurs (en particulier pour améliorer l'accès aux dictionnaires), analyseur grammaticaux (entre autres pour la vérification grammaticale), analyseurs stylistiques (pour la critique de style), etc. La disponibilité de corpus bilingues peut également permettre de réaliser, en amont des traducteurs et rédacteurs, des travaux qui soient réutilisables par eux : génération de dictionnaires bilingues, tables statistiques, etc. Par ailleurs, la communication avec d'autres spécialistes doit pouvoir se faire aisément : communication d'une liste de mots nouveaux au terminologue, questions au rédacteur, etc.

Conclusion

Vu l'évolution actuelle de la technologie, il n'est pas utopique d'envisager un poste de travail intégré et configurable, fournissant de nombreux outils d'aide à l'utilisateur. Toutes les activités présentées dans cet article sont en fait très liées, et l'expérience montre qu'elles doivent être considérées dans leur ensemble.

L'utilisation du TALN et des outils annexes dans les activités de gestion documentaire est devenue un axe stratégique pour de nombreuses entreprises. Pour SITE, la maîtrise d'œuvre du projet EUROLANG et le développement du poste d'enrichissement lexicographique pour le CNET s'inscrivent tout à fait dans cet axe.

Bibliographie

BACHUT D. (1991) « Industrialisation d'un système de TAO français-anglais dans le domaine technique », *Colloque Génie linguistique 91*, Versailles, 16-17 janvier 1991.

BOITET C., *et al.* (1982) « ARIANE-78 : an Integrated Environment for Automated Translation and Human Revision », *Coling 82*, Prague.

HEYER G. *et al.* (1991) « Motivation, Goals and Milestones of Esprit II project MULTIPLEX », *Colloque Génie linguistique 91*, Versailles, 16-17 janvier 1991.

MARET D. (1991) « Poste de travail lexicographique générique du CNET », proposé au *Colloque d'Avignon 91*.

NOSSIN M. (1991) « Le projet GENELEX : EUREKA pour les dictionnaires électroniques », *Colloque Génie linguistique 91*, Versailles, 16-17 janvier 1991.

VAUQUOIS B. et BOITET C. (1985) « Automated Translation at Grenoble University », *Computational Linguistics*, vol. 11, n° 1, January-March 1985.