INDEXAGE LEXICAL AU GETA

COLING-86
Bonn

Jedrzej BUKOWSKI

GETA, BP 68
Universite Scientifique et Médicale de Grenoble
38402 Saint-Martin-d'Heres, FRANCE

RESUME: L'aspect lexicographique de la traduction assistee par ordinateur est présenté et illustré par des exemples de traduction du russe en français realisée par le GETA a Grenoble.

MOTS-CLES: TAO (traduction assistee par ordinateur), analyse morphologique, transfert lexical, generation morphologique, indexage.

DEFINITIONS:

ARIANE - systeme informatique qui donne aux linguistes la possibilite de réaliser un environnement interactif et des metalangues specialisees (LSPL) dans lesquelles ils ecrivent les données linguistiques pour construire une application TAO et realiser des traductions.

SYSTEME - système de TAO: il se compose d'ARIANE et des donnees linguistiques ("linguiciel").

FORME (OCCURRENCE CHAINE) - suite de caracteres entre deux blancs. Il faut noter que les notions "mot", "forme" ou encore "occurrence" suggerent qu'il s'agit d'un seul élement. Or, dans des cas frequents, il peut être question de "tournures" : il y a "tournure" chaque fois qu'une signification est exprimée par plusieurs chaînes de caractères et que la presence d'une chaîne change le sens d'une autre chaîne.

BASE - racine communement admise permettant de produire une ou plusieurs formes d'un lemme à l'intérieur d'un paradigme donné.

LEMME - ensemble de formes d'un mot à l'intérieur d'un paradigme communément admis; il est noté par une forme representative dite canonique.

FORMATS ET PROCEDURES - Chaînes de caractères representant, sous formes codées dans les dictionnaires de chaque etape de la traduction, les informations lexicales relatives au comportement des mots, aux classes syntaxiques, dérivationnelles et sémantiques, et aux phenomenes contrastifs.

UNITE LEXICALE (UL) - ensemble arbitraire de lemmes qui constituent une communauté de signification. Une UL possede une structure lemmatique determinée, qui contient les types des lemmes consideres par l'indexeur comme ses dérivés.

VARIABLES - ensembles d'informations linguistiques et "stratégiques" permettant de décrire le mot indexé et de preciser les paramètres permettant d'assurer un meilleur fonctionnement du système. Elles sont définies par leurs noms et par l'ensemble de leurs valeurs. Les variables sont predéfinies et déclarees soit sous forme d'expressions booleennes pour chacune des variables individuelles, soit de combinaisons des variables.

DICTIONNAIRES - Ensembles d'articles qui contiennent, dans chaque etape de la traduction, les informations nécessaires pour reconnaître le mot du texte source et identifier son comportement, choisir son equivalent en langue cible et préciser les informations nécessaires correspondantes à son comportement. Certains dictionnaires d'analyse morphologique, notamment celui des paradigmes et, dans une moindre mesure, celui des prefixes, sont prédéfinis et declarés au système.

I. INTRODUCTION

Différents logiciels et systemes de traitement de texte mis au point par les informaticiens du GETA, ainsi que les bases théoriques du système du GETA ont ete traités dans la littérature (voir notamment 1,2,3,4 et 5). Dans ce qui suit nous nous concentrerons sur la creation et la maintenance des dictionnaires dans les phases consécutives de la traduction.

I. TRAVAIL LEXICOGRAPHIQUE DANS LA TRADUCTION

Pour qu'un texte source puisse être traduit en langue cible, il faut reconnaître les mots de ce texte, en comprendre le contenu, predefinir une structure et les élements lexicaux équivalents en langue cible, et produire le texte correspondant en langue cible.

Le processus de traduction est ainsi divisé en trois etapes successives, analyse, transfert et génération, chacune contenant une phase structurale et une phase lexicale. Pour ce qui nous intéresse, nous allons nous concentrer sur les phases lexicales du processus: analyse morphologique (AM), transfert lexical (TL) et génération morphologique (GM).

Parmi les éléments lexicaux les dictionnaires des bases de l'analyse morphologique constituent un ensemble pratiquement infini, et vehiculent des informations variees en fonction du domaine que décrit le texte source et souvent du contexte à l'intérieur même d'un texte.

Le travail lexicographique consiste à tenir compte de cette richesse et de cette variété. Il est réalisé par les indexeurs, qui disposent du système ARIANE, à l'intérieur duquel ils creent et entretiennent leurs dictionnaires. Les indexeurs indexent les mots du texte source dans les dictionnaires propres à chaque étape "lexicale" de la TAO.

II. INDEXAGE: DEFINITION ET PRINCIPES

L'indexage est un ensemble d'opérations qui permettent de realiser les dictionnaires de chacune des trois phases de la traduction, soit d'attribuer aux mots du texte source les informations décrivant leur comportement linguistique et syntaxique, de choisir leurs équivalents et de preciser les informations relatives à leur comportement dans la langue cible.

Dans le dictionnaire AM, l'indexeur définit, à partir d'un mot du texte source, sa base, les formats morphologique et syntaxique qui décrivent son comportement, et une UL qui représente la structure lemmatique propre à ce mot. Cette dernière constitue l'entrée du dictionnaire TL, dans lequel l'indexeur choisit l'équivalent en langue cible représenté par une UL, et tient compte de différents lemmes appartenant à la structure lemmatique de l'UL source. L'UL cible du dictionnaire TL permet d'accéder au dictionnaire GM. Dans ce dernier l'indexeur aboutit à une ou à plusieurs bases qui correspondent à la structure lemmatique de la langue cible.

L'organisation des données lexicales et des dictionnaires repond à deux principes importants. Le premier, c'est celui de la cohérence des dictionnaires. Toute la structure lemmatique définie en AM doit être reprise dans les deux autres étapes, ce qui veut dire, que non seulement toutes les UL source définies en AM doivent être reprises en TL et en GM, mais que l'indexeur doit choisir les formats et les procédures dans tous les dictionnaires de façon à rendre compte de toute les formes appartenant à l'UL retenue.

Le second principe est celui de l'économie de place. En effet, c'est de la définition de la structure lemmatique de l'UL que depend le nombre de formes à indexer, et par conséquent le volume occupé dans les dictionnaires. C'est pourquoi, les indexeurs essayent d'entrer le maximum de lemmes dans une seule UL, ce qui demande, au fur et à mesure des nouveaux textes à traduire, une mise à jour des dictionnaires.

Les dictionnaires du système GETA sont écrits selon une syntaxe très precise et l'indexeur suit un schema d'indexage symmétrique:

```
+-----------------------------------------------------+
!      BASE + UL (source) ----> UL + BASE (cible).    !
+-----------------------------------------------------+
```

Remarquons que les dictionnaires AM et GM sont monolingues et "déterministes", puisque l'indexeur y code les informations propres aux langues source et cible, tandis que le dictionnaire TL est bilingue et l'indexeur décide du choix d'équivalent et de tel ou tel élément dérivationnel ou contrastif qu'il juge necessaire.

Ces dictionnaires sont écrits dans des LSPL (langages specialisés pour la programmation linguistique), puis compiles dans une representation interne. Pendant l'execution, ils resident en memoire virtuelle.

III. BREF RAPPEL DES OPERATIONS LINGUISTIQUES

1. ANALYSE MORPHOLOGIQUE

L'entrée est le texte source, dans la transcription interne, avec les marques de pré-edition, des ordres de formatage SCRIPT et des occurrences spéciales pour les "hors-textes" (figures, tableaux, etc...). La sortie est une structure arborescente décorée (ou annotée). La racine identifie le texte, sa décoration contient ULTXT comme valeur de la variable UL. Les noeuds fils directs de la racine correspondent aux phrases, et contiennent la valeur ULPRA (pour "phrase"). Les noeuds au niveau 2, sous les noeuds de la phrase, correspondent aux occurrences et contiennent la valeur ULOCC. Les noeuds de niveau 3 correspondent aux resultats de l'analyse morphologique (un ou plusieurs pour chaque occurrence). Pour un mot composé le noeud de ce niveau contient l'unite lexicale ULMCP et domine les noeuds contenant les résultats de l'analyse morphologique des différents composants. Le LSPL de cette etape est ATEF; il transforme les chaînes en arbres.

2. TRANSFERT LEXICAL

Au cours de cette étape, aux UL source sont attribuées les UL cible. Le dictionnaire de TL est ecrit en LSPL TRANSF qui transforme les arbres en arbres. Chaque noeud de l'arbre d'entree est remplacé par un sous-arbre dans l'arbre de sortie, choisi entre plusieurs possibilites en fonction de l'évaluation d'un prédicat sur les attributs du noeud d'entrée et de ses voisins immédiats. Dans le cas simple, le sous-arbre est réduit à un seul noeud. Dans des cas compliqués, le sous-arbre choisi peut soit donner plusieurs équivalents possibles, parmi lesquels le choix pourrait s'opérer dans l'étape suivante, soit predire une tournure en langue source (voir le chapitre V et le doc. n° 9).

3. GENERATION MORPHOLOGIQUE

La suite (lue de gauche à droite) des feuilles de l'arbre résultant de la génération syntaxique est l'entree de la GM, écrite en LSPL SYGMOR qui transforme les arbres en chaînes. La grammaire dirige la construction des occurrences du texte cible en testant les valeurs des variables du noeud courant (et d'un contexte borné), et en se référant aux dictionnaires des chaînes accedes par l'UL ou à d'autres variables pour obtenir les divers morphes à combiner (bases, préfixes, affixes, etc...).

IV. FORMES D'ARTICLES DES DICTIONNAIRES

Dans chacune des étapes, les dictionnaires ont une forme particuliere dont voici les schemas suivis d'exemples. L'indexage est réalisé à l'aide des "guides d'indexage" (voir les documents 5, 7, et 8). En répondant aux questions consécutives qui tiennent compte des phénomenes linguistiques, l'indexeur précise ceux qui concernent le mot indexé et arrive aux codes qu'il introduit dans ses dictionnaires. Les éléments soulignés sont obligatoires dans chaque dictionnaire.

1. EN ANALYSE MORPHOLOGIQUE

Dans cette étape il y a les cinq dictionnaires suivants: des préfixes, des paradigmes, des bases, des tournures, et un dictionnaire en réserve.

```
+-----------------------------------------------------+
!BASE          ==FM          (FS   ,UL(russe)      ).  !
!                                                     !
!METAN         ==B14B        (N4B  ,METAN          ).  !
+-----------------------------------------------------+
```

2. EN TRANSFERT LEXICAL

En transfert lexical il y a sept dictionnaires "thématiques", dans lesquels on introduit les équivalents en fonction du contexte ou domaine traité dans un texte. Il y a ainsi un dictionnaire contenant des termes de vocabulaire de base, un pour la technique générale, un pour les mathematiques et la physique, un pour la chimie et la métallurgie, un pour la navigation, l'aviation et l'astronautique, un pour les sciences sociales, la géographie et la politique, et un pour les termes généraux complémentaires. Ces différents dictionnaires sont numérotés, ce qui permet de réaliser une certaine adaptation au domaine en choisissant un ordre de priorité correspondant au texte traduit.

```
+-----------------------------------------------------+
! 'UL(russe)'    ==PCP  //UL(française)  ,*FAF ,PAF   !
!                                                     !
! 'METAN'        ==     //'ME!1THANE'    ,*NUL ,$G1.  !
+-----------------------------------------------------+
```

3. EN GENERATION MORPHOLOGIQUE

Dans cette étape il y a un seul dictionnaire:

```
+-----------------------------------------------------+
! UL(française)     ==PRC     /FM      /BASE.         !
!                                                     !
! ME!1THANE         ==        /MOT     /ME!1THANE.    !
+-----------------------------------------------------+
```

Dans les schémas ci-dessus, nous avons montré la forme générale de chaque dictionnaire et un exemple concret, celui du mot russe "METAN". Le comportement morphologique de ce mot ne presente aucune particularite et aucune dérivation n'est possible, ce qui fait que son indexage est particulierement simple.

V. CAS PLUS COMPLEXES

Il arrive souvent, que l'indexage est bien plus complexe. Par exemple, le verbe imperfectif russe KRUTITQ (tordre, vriller) donne, dans le dictionnaire AM, les quatre bases suivantes: KRUKH (avec le FM B64A3I, pour la 1ère personne du présent de l'indicatif), KRUT (avec le FM B65E0I, pour les formes de l'infinitif, l'imperatif et pour le reste du paradigme verbal), KRUT (avec le FM B20D1, pour le substantif féminin KRUTKA) et KRUTILQN (avec le FM B80B, pour l'adjectif KRUTILQNYIJ dérivé de la forme infinitive). A chacune de ces bases l'indexeur attribue aussi les FS correspondants et une même UL:

```
+-----------------------------------------------------+
!KRUKH         ==B64A3I  (U27      KRUTITQ          ).  !
!KRUT          ==B65E0I  (U27      KRUTITQ          ).  !
!KRUT          ==B20D1   (XZ       KRUTITQ          ).  !
!KRUTILQN      ==B80B    (M        KRUTITQ          ).  !
+-----------------------------------------------------+
```

Dans le dictionnaire TL, à partir de cette UL, l'indexeur rend compte de la dérivation adjectivale: la PCP "ADJVRB", les PAF "GDP" et "B1C" et l'UL française

equivalente TORDRE, permettent d'obtenir la forme DE TORSION. Dans le même article, l'indexeur doit aussi rendre compte de la traduction courante du verbe source. Ainsi, les conventions de syntaxe permettent, après la ou les lignes correspondant aux PCP, d'introduire une ligne (ou plus) avec cette traduction. Dans le cas courant, il y a dans l'ordre les UL françaises TORDRE et VRILLER, une PAF spéciale, "XL", spécifiant que la dernière est un synonyme de la précédente.

Le dictionnaire de TL permet aussi de rendre compte du contexte, c'est-à-dire d'indexer les tournures (voir DEFINITIONS). Les détails d'indexage des tournures sont donnés dans le document (9). Disons seulement que l'indexage des tournures suit un schéma spécifique: l'indexeur indique d'abord, sur le sommet principal, l'equivalent courant du mot origine, et ensuite le contexte origine (par un sommet '<CONTEXTE>' avec le FM "CHOIX" et sans aucune PAF (c'est le seul cas d'absence de PAF en dictionnaire TL) précisant ce contexte par les UL cible équivalentes, le sommet '<CONTEXTE>' et les sommets des UL cible étant les dépendants du sommet principal.

Ainsi avec le verbe KRUTITQ, l'apparition dans un nouveau texte russe du contexte MASHINA ("machine") a imposé le changement de son équivalent: en effet, en russe KRUTILQNAYA MASHINA ("krutilqnaya" est un adjectif dérivé du verbe "krutitq") signifie RETORDEUSE, ce qui est exprimé dans le dictionnaire de la façon suivante:

```
+-------------------------------------------------+
!'KRUTITQ'  ==$AUJVRB / 0(C(1,E))/   2             !
!                     0:'TORDRE'     2+NUL ,$GPD,$B1C;!
!                     C:'<CONTEXTE>' 2*CHOIX;       !
!                     1:'MACHINE'    2+PERE,$EFF;   !
!                     E:'RETORDRE'   2+VZI ,$AZ/    !
!                     / 0(1)/        2              !
!                     0:'TORDRE'     2+VZI ,$VF,$R1; !
!                     1:'VRILLER'    2+VZI ,$VM,$XL. !
+-------------------------------------------------+
```

On voit que le contexte russe n'apparaît dans le dictionnaire TL que sous forme de son equivalent en langue cible. De plus, la PAF "EFF" permet de ne pas le reproduire dans l'equivalent contextuel RETORDEUSE obtenu grâce à la PAF "AZ". Pour TORDRE les PAF "VF" et "R1" correspondent, respectivement, à l'existence du substantif verbal féminin TORSION dérivé de l'infinitif, et au phénomène contrastif de réflexivité, possible pour le verbe russe et pour le verbe français. Pour VRILLER, la "VM" dénote le substantif verbal masculin VRILLAGE, et la "XL" le fait que les deux verbes français sont synonymes.

Toutes les UL françaises doivent naturellement être indexées dans le dictionnaire GM. Pour TORDRE, il a y par exemple: la base TORDANT, avec la PRC "AD" (pour l'adjectif) et le FM "NOIR" (pour les formes TORDANT (0, S, E, ES)), la base TORDEU, avec la PRC "ND" (pour le substantif dérivé du verbe) et le FM "EUSE" (pour les formes TORDEU (K, RS, SE, SES), la base TORDAGE, avec la PRC "NVBMAS" (pour le nom masculin dérivé du verbe) avec le FM "MOT" (pour les formes TORDAGE (0, S), la base TORSION, avec la PRC "NVBFEM" (pour le nom féminin dérivé du verbe) avec le FM "MOT" (pour les formes TORSION (0, S) et enfin, une dernière ligne de ce dictionnaire, obligatoirement sans PRC, donne la base TORD, à partir de laquelle on peut générer tout le paradigme du verbe TORDRE.

Ceci donne, les articles suivants:

```
+-------------------------------------------------+
!<CONTEXTE>          ==         /RIEN    /.       !
!                                                 !
!MACHINE             ==         /MOT     /MACHINE. !
!                                                 !
!RETORDRE            ==ND       /EUSE    /RETORDEU, !
!                    ==NVBMAS   /MOT     /RETORDAGE, !
!                    ==         /RENDRE  /RETORDRE. !
!                                                 !
!TORDRE              ==AD       /NOIR    /TORDANT, !
!                    ==ND       /EUSE    /TORDEU,  !
!                    ==NVBFEM   /MOT     /TORSION, !
!                    ==NVBMAS   /MOT     /TORDAGE, !
!                    ==         /RENDRE  /TORD.    !
!                                                 !
!VRILLER             ==NVMFEM   /MOT     /VRILLAGE, !
!                    ==         /MOT     /VRILL.   !
+-------------------------------------------------+
```

VI. REMARQUES FINALES

Les dictionnaires de l'equipe russe de GETA atteignent actuellement des dimensions importantes (10 000 lignes en moyenne dans les trois étapes, ce qui correspond approximativement à quelquess 7000 ou 8000 UL differentes). En dehors des problèmes de l'indexage, parmi lesquels les choix terminologiques appartiennent aux plus difficiles, l'indexeur est confronté avec les nouveaux textes à traduire, ce qui implique notamment la révision constante des equivalents déjà attribués, et ceci soit à cause des nouveaux contextes, soit de la spécialité croissante des textes, les deux obligeant aux raffinements.

Du point de vue informatique certains outils complémentaires ont été implémentés, et notamment le logiciel ATLAS qui offre une aide automatisée à l'indexage. Il permet à l'indexeur de manipuler sur écran les fichiers des formats et des procédures et de remplir automatiquement ses dictionnaires. L'indexeur peut aussi utiliser un dictionnaire terminologique qu'il appelle sur son écran et dont il peut automatiquement transferer les mots dans le dictionnaire indexé. Mais ceci ne peut qu'être une aide appreciable. L'essentiel du travail reside dans les connaissances linguistiques et terminologiques de l'indexeur et de la souplesse du système dont les contraintes informatiques sont des sources possibles de problèmes.

BIBLIOGRAPHIE

1.B.VAUQUOIS (1979),
"Aspects of automatic translation in 1979",
IBM-Japan, Scientific Program, July 1979.

2.Ch.BOITET, N.NEDOBEJKINE (1981),
"Recent developments in Russian-French Machine at Grenoble",
Linguistics 19, 199-271 (1981)

3.Ch.BOITET, P.GUILLAUME, M.QUEZEL-AMBRUNAZ (1982)
"ARIANE-78: an integrated environment for automated translation and human revision",
Proceedings COLLING 82, North-Holland, Linguistic Series N° 47, 19-27, Prague, July 1982.

4.Ch.BOITET (1985),
"Traduction (assistée) par ordinateur: ingenierie logicielle et linguicielle",
Proceedings Congrès AFCET, Grenoble, novembre 1985, vol.1.

5.Ch.BOITET, P.GUILLAUME, M.QUEZEL-AMBRUNAZ (1985)
"ARIANE-85: a case study in software evolution from Ariane-78.4 to Ariane-85",
Seminar on MT, Hamilton, Colgate University, 14 - 16 August 1985.

6.N.NEDOBEJKINE (1985)
"Croissance de la base lexicale du système de TAO russe-français, version 85-06: Guide d'indexage pour l'analyse",
GETA, R.R. DRET n°70, Grenoble, Juillet 1985.
14 - 16 August 1985.

7.N.NEDOBEJKINE (1985)
"Croissance de la base lexicale du système de TAO russe-français, version 85-06: Guide d'indexage pour le transfert",
GETA, R.R. DRET n°71, Grenoble, Juillet 1985.
14 - 16 August 1985.

8.N.NEDOBEJKINE (1985)
"Croissance de la base lexicale du système de TAO russe-français, version 85-06: Guide d'indexage pour la generation",
GETA, R.R. DRET n°72, Grenoble, Juillet 1985.
14 - 16 August 1985.

9.A.BUKOWSKI, C.TORRE (1986)
"Tournures lexicales non-figées en transfert"
GETA, R.R. DRET n°83, Grenoble, Mars 1986.