Variables et catégories grammaticales dans un modèle Ariane

paper submitted for
COLING 86
Bonn

by

Jean-Philippe GUILBAUD

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

(April 1986)

RESUME

Toutes les catégories grammaticales utilisées dans un modèle de traduction Ariane sont formalisées et codées de façon mnémonique en tant que variables et valeurs de variables. L'ensemble des variables d'un modèle donné constitue le vocabulaire du métalangage qui permet de décrire la langue source et la langue cible de ce modèle.

La structure de données du système est une arborescence dont chaque noeud porte une décoration. Les décorations contiennent les variables déclarées pour le système et affectées de certaines valeurs. Les variables apparaissent également dans les grammaires d'analyse, de transfert et de génération, dans les dictionnaires monolingues d'analyse ou de génération et bilingues de transfert lexical, ainsi que dans les spécifications de modèle linguistique (grammaires statiques).

MOTS CLEFS

Variables, valeurs de variables, unité lexicale, lemme, mot, arborescence, décoration, grammaire statique, analyse morphologique (AM), analyse structurale (AS), transfert lexical (TL), génération syntaxique (GS), génération morphologique (GM), structure linguistique interface (SLI), source, cible.

I. INTRODUCTION

Chaque concept grammatical est une classe qui regroupe un certain nombre d'éléments possibles. Le linguiste est libre d'organiser ses classes comme il l'entend et de donner l'interprétation qu'il veut aux diverses combinaisons d'éléments. Par exemple

- Genre (masculin, féminin, neutre),
- Nombre (singulier, pluriel),
- Relation (hypotaxe, parataxe, prostaxe),
- Elément logique (thème, rhème, phème),
- Domaine (général, médecine, aviation),

sous forme codée, reçoivent un nom et un ensemle de valeurs, librement choisis par le linguiste, par exemple:

- GNR(m,f,n);
- NBR(sin,plu);
- RELAT(hypo,para,pro);
- CLOG(th,rh,ph);
- DOMAINE(gen,med,avion).

A titre d'exemple, dans le cas du nombre, on peut désigner la classe, un de ses éléments, ou aucun d'eux :

- NBR(sin,plu);
- NBR(plu);
- NBR(Ø).

Lorsqu'on a établi la liste du vocabulaire du métalangage de description d'une langue, il est possible de définir des conventions d'interprétation, par exemple:

- A(1,2) ==> 1 et 2;
- B(X,Y) ==> X ou Y;
- P(T,G) ==> NBR(Ø).

II. LES VARIABLES DANS LA SLI SOURCE

La structure linguistique interface (SLI) de la langue source est une arborescence décorée qui rend compte de la structure de chacune des phrases de l'énoncé à traduire. Sa géométrie définit un parenthésage de groupes, tandis que les variables reflètent la sémantique de ce parenthésage.

1. STRUCTURE SYNTAXIQUE DE SURFACE

La variable <classe syntagmatique> attribue un nom à chacun des groupes imbriqués de la structure (groupe nominal, groupe adjectival, subordonnée infinitive, etc.).

La variable <fonction syntaxique> sert à indiquer le rôle syntaxique de chaque groupe (sujet, objet, attribut du sujet, attribut de l'objet, etc.) par rapport au noyau de ce groupe (gouverneur), et à identifier ce dernier.

2. STRUCTURE SYNTAXIQUE DE TRANSFERT

Cette structure constitue le "niveau profond" qui demeure inchangé de la langue source à la langue cible, à cette restriction près que certaines inversions dans la numérotation des arguments peuvent devoir être effectuées pour la cible.

La variable <relation logique> permet d'identifier en les numérotant les arguments de tous les gouverneurs de groupe prédicatifs. La numérotation est invariante par rapport à la classe syntagmatique et la fonction syntaxique des groupes.

La variable <relation sémantique> permet de spécifier la nature sémantique des relations logiques ainsi que la fonction sémantique des groupes non-arguments. <relation logique> et <relation sémantique> forment un tout.

Les quatre variables présentées ci-dessus se complètent et rendent compte de trois niveaux d'interprétation au moyen d'une seule géométrie: La SLI est en quelque sorte une structure génératrice de trois structures différentes.

3. REPRESENTATION LEXICALE

La représentation lexicale des mots et tournures des phrases se fait à trois niveau:

La variable UL (unité lexicale) permet de désigner une classe particulière de lemmes qui renvoient au même concept ou archilexème de référence et dérivent les uns des autres. Par exemple, UL(construire) pourra désigner les lemmes: construire, construction, constructeur, constructible, constructibilité, constructif, constructivité; et UL(manger), les lemmes: manger, mangeur, comestible, comestibilité.

Chaque UL possède un lemme particulier à partir duquel on dérive tous les autres. On l'appelle lemme origine et il sert à dénoter l'UL.

Les variables de <catégorie> et de <dérivation> permettent ensuite de dénoter n'importe quel sous-ensemble de lemmes.

4. ACTUALISATION MORPHOLOGIQUE

Les variables d'actualisation morphologique telles que <nombre>, <genre>, <personne>, <temps>, <mode>, etc. décrivent avec précision le mot du lemme trouvé dans le texte. Par exemple:

- Ul(construire)
- <catégorie>(nom)
- <dérivation>(verbe vers le nom d'agent)
- <genre>(masculin)
- <nombre>(pluriel)

est une façon de représenter le mot constructeurs.

5. ACTUALISATION SYNTAXIQUE

Les variables d'actualisation syntaxique décrivent le type d'énonciation de la phrase (interrogatif, exclamatif, conditionnel, etc)) et sa réalisation verbale (passif, actif, progressif, etc.).

III. LES VARIABLES DANS LE DICTIONNAIRE D'ANALYSE

Dans le système Ariane les dictionnaires d'analyse ne sont ouverts qu'une seule fois: pendant l'analyse morphologique. Leur fonction est double, permettre le découpage des mots en AM et assigner aux mots toutes les informations utiles au bon déroulement de l'AS.

Ces dictionnaires sont des dictionnaires de morphes. Certains contiennent des bases lexicales, d'autres divers affixes flexionnels et dérivationnels.

Les variables dans les dictionnaires informent d'une part sur la référence lexicale, les paradigmes de flexion, dérivation et conjugaison, d'autre part sur les propriétés distributionnelles du point de vue de la puissance et de la valence. Les variables paradigmatiques permettent d'analyser correctement chaque mot du texte en vue d'identifier lemme et UL de référence. Les variables distributionnelles permettront au moment de l'analyse structurale ultérieure d'opérer les bons parenthésages de groupes.

Valence et puissance sont généralement exprimées par des valeurs de

- catégorie morphosyntaxique ou syntagmatique;
- cas de déclinaison nominale (accusatif, etc.);
- prépositions régies;
- traits sémantiques.

Caractérisation et distribution sémantiques des lexèmes connaissent les limites qui sont celles de la sémantique de traits. En général, on utilise un nombre restreint de traits, afin d'éviter le danger toujours imminent des contradictions et incohérences. Il s'agit uniquement de compléter les traits grammaticaux si ceux-ci s'avèrent par trop insuffisants. On n'ignore pas, par exemple, que tout ce qui relève du concret peut servir à formuler de l'abstrait et que toute abstraction peut devenir réalité désignable: il y a des échanges permanents entre le rhème et le thème (cf. Zemb). Pour ne parler que de l'opposition concret-abstrait...

IV. LES VARIABLES DANS LA SLI CIBLE

La structure interface cible est l'entrée du processus de génération syntaxique de la langue cible. Elle va être transformée progressivement jusqu'à obtenir la structure de surface des phrases du texte traduit.

La SLI cible est l'image de la SLI source. Les variables contenues dans ses décorations sont la traduction en cible du vocabulaire du métalangage de description de la source. Les références lexicales sont nouvelles, le niveau profond des relations logiques et sémantiques restant inchangé. Les niveaux moins profonds ont été automatiquement traduits, mais doivent être tous recalculés et réajustés en fonction

- des nouvelles propriétés distributionnelles des nouvelles ULs issues du transfert lexical;
- des caractéristiques de la grammaire de la langue cible, qui ne sont évidemment pas celles de la grammaire de la langue source.

Ainsi, on fixe la valeur de fonction syntaxique et celle de classe syntagmatique d'un groupe à partir de sa valeur de relation logique et des variables distributionnelles portées par le gouverneur du groupe dont il dépend. Pour citer un exemple de transformation effectuée sur la SLI cible pendant le traitement GS.

En fin de GS tout a été vérifié ou recalculé. Les feuilles de l'arbre qui portent une référence lexicale sont dans le bon ordre (celui des mots de la phrase) et affectées des bonnes valeurs de catégorie, de dérivation et d'actualisation morphologique, de sorte que l'automate de génération morphologique peut produire les mots du texte traduit en deux étapes:

- calculer la base lexicale du bon lemme de l'UL;
- calculer les affixes qui font le bon mot du lemme de l'UL.

V. LES VARIABLES DANS LE DICTIONNAIRE DE TRANSFERT LEXICAL

En TL, il y a tout d'abord mise en correspondance d'une liste de valeurs d'ULs sources avec une liste de valeurs d'ULs cibles. Chaque UL cible est accompagnée d'un ensemble de valeurs de variables relatives à sa distribution syntaxique. Ce sont ces nouvelles propriétés distributionnelles qui seront à la base du calcul, dans la SLI cible, des catégorie syntagmatiques et fonctions syntaxiques des éventuels groupes dépendants.

Les conditions de choix entre plusieurs cibles pour une source sont libellées à l'aide de valeurs de variables: celles qui sont contenues dans les décorations de la SLI source. Il est rare qu'une valeur d'UL source suffise à elle seule pour déterminer l'équivalent cible. Très souvent il faut tester la distribution effective et parfois même l'actualisation.

VI. LES VARIABLES DANS LE DICTIONNAIRE DE GM

Les dictionnaires de GM sont des dictionnaires monolingues qui contiennent la liste exhaustive des morphes constitutifs des mots de la cible. Comme en AM, il y a des dictionnaires de bases lexicales et des dictionnaires d'affixes grammaticaux.

Chaque article de dictionnaire est une valeur ou un ensemble de valeurs de variables auquel est affecté un morphe (une chaîne de caractères) et éventuellement des valeurs de variables paradigmatiques qui permettent, pour un morphe lexical, de trouver préfixe, suffixe et/ou désinence, dans un dictionnaire d'affixes, en fonction des valeurs de variables dérivationnelles et d'actualisation issues de GS.

A l'inverse de l'AM où l'on va des morphes de la langue vers les valeurs de variables par le biais de dictionnaires et de grammaires, la GM va des variables aux morphes de la langue par l'intermédiaire de dictionnaires et d'une grammaire.

VII. VARIABLES DYNAMIQUES VS VARIABLES STATIQUES

On appelle variables dynamiques celles d'un modèle exécutable en Ariane. Les variables statiques sont celles que l'on utilise dans les grammaires statiques.

Les spécifications linguistiques d'un modèle se font au moyen de grammaires statiques. Ces grammaires décrivent les langues source et cible d'un modèle en utilisant une autre structure de données que le système Ariane mais utilisent la même notation, sous forme de variables, des catégories grammaticales. Elles restent indépendantes de toute stratégie d'approche dynamique (stratégie d'analyse ou de génération).

Avant même que d'écrire de telles grammaires, on dresse la liste de toutes les notions utiles à la description de la langue dont on veut faire l'analyse ou la synthèse. Elles sont ensuite codées sous forme de variables et de valeurs de variables.

VIII. QUELQUES VARIABLES DE L'ALLEMAND

Les variables de l'allemand couvrent l'ensemble des notions linguistiques relatives

- aux catégories morphosyntaxiques
- aux dérivations morphosyntaxiques
- aux cas de flexion
- à l'actualisation morphologique
- à l'actualisation syntaxique
- au degré de l'adjectif
- aux propriétés syntaxiques
- aux propriétés sémantiques
- à la distribution syntaxique et sémantique
- aux classes syntagmatiques
- et aux fonctions syntaxiques.

1. LES PARTIES DU DISCOURS EN ALLEMAND

Dans le modèle allemand, on décrit les parties du discours à l'aide d'une variable de catégorie et plusieurs variables de sous-catégories.

L'ensemble des valeurs ne constitue pas un tout homogène, pas plus que les parties du discours n'en font un dans la plupart des grammaires des langues du monde. Les critères de distinction peuvent être d'ordre morphologique, syntaxique, sémantique ou logique, et dans tous les cas calqués sur la plupart des grammaires traditionnelles.

On a jugé utile d'ajouter à la liste traditionnelle les valeurs de

- prédicateur (ces "adverbes de phrase" dont la fonction est d'établir la liaison entre thème et rhème - cf Zemb (1978));
- signe d'édition;
- forme non-alphabétique (pour les équations mathématiques et autres formules chimiques);
- de référence (de pièces détachées, par exemple).

La sous-catégorisation de l'adjoint (adjectif ou adverbe) est réalisée en fonction de la puissance de ce dernier: un adjoint peut déterminer

- un verbe attributif;
- un verbe non-attributif;
- un nom;
- un autre adjoint;
- un sous-ensemble quelconque des puissances précédentes.

2. L'EXPRESSION DE LA SUITE FLEXIONNELLE COHERENTE

La cohérence d'une suite flexionnelle telle que

dIE grossEN TierE

s'exprime par rapport au cas, au genre grammatical, au nombre et au type de déclinaison (fort/faible). Etant donné que ces catégories sont liées dans toute suite flexionnelle, on a décidé de créer les 8 variables suivantes:

- cas singulier masculin fort;
- cas singulier masculin faible;
- cas singulier féminin fort;
- cas singulier féminin faible;
- cas singulier neutre fort;
- cas singulier neutre faible;
- cas pluriel fort;
- cas pluriel faible.

qui prennent toutes leur valeur dans l'ensemble unique

(nominatif,accusatif,datif,génitif).

Toute désinence d'article ou d'adjectif n'est forte ou faible que par rapport à certains cas. Un substantif au pluriel est toujours fort. Un substantif au singulier ne l'est que s'il relève du non-discret.

Dans la grammaire d'analyse, on fait prévaloir les règles suivantes:

- FORT + FORT --> FORT;
- FORT + FAIBLE --> FORT;
- FAIBLE + FORT --> FAIBLE;
- FAIBLE + FAIBLE --> FAIBLE.

Une suite cohérente nominale doit avoir une résultante forte et l'intersection des valeurs portées par ses éléments ne doit pas être vide.

3. LES VALENCES

Les valences d'un lexème sont exprimées au moyen de trois variables principales (VALO, VAL1, VAL2) : Elles spécifient le type syntagmatique des arguments 0,1 et 2 exigés par ce lexème.

Un argument peut être

- un groupe nominal;
- un groupe prépositionnel;
- une subordonnée complétive;
- une infinitive;
- ou une interrogative indirecte.

Si l'argument est un groupe nominal (prépositionnel ou non) les variables de valence indiquent le cas (et la préposition) régie.

Exemple: BEMALEN -> VAL1(GN), VAL2(GP), GN1(ACC), GP2(MIT).

BIBLIOGRAPHIE

CHAPPUY, S. (1983). Formalisation de la description des niveaux d'interprétation des langues naturelles. Etude menée en vue de L'analyse et de la génération au moyen de transducteurs. Thèse de 3ème cycle, INPG Grenoble.

ZEMB, J.-M. (1978). Vergleichende Grammatik Französisch-Deutsch Teil 1. Mannheim: Bibliographisches Institut. Duden-Sonderreihe vergleichende Grammatik; Bd 1.

VAUQUOIS, B. et CHAPPUY S. (1985). Static Grammars: a formalism for the description of linguistic models. Proceedings of the Conference on Theoretical and Methodological issues in Machine Translation of Natural Languages, Colgate University, Hamilton, NY, USA, August 14-16, 1986.

STAHL, G. (1984). L'analyse syntaxique automatique de l'allemand. CNRS, Paris.

-o-o-o-o-o-o-o-o-o-