Une représentation sémantique et un système de transfert pour une traduction de haute qualité.

Présentation de projet
Avec démonstration sur machines SUN

K. BOESEFELDT et P. BOUILLON
ISSCO
54, route des ACACIAS
1227 GENEVE
kathy@divsun.unige.ch/pb@divsun.unige.ch

1. Introduction

Le projet que nous présentons dans cet article a pour but l'implémentation d'un système de traduction automatique pour des bulletins d'avalanches, qui utilise ELU, un environnement linguistique d'unification, basé sur la méthode du transfert. Un tel projet nécessite l'écriture et la maintenance de plusieurs grammaires et d'un ensemble de règles de transfert.

Dans cet article, nous commençons par présenter de manière très succincte le projet. Nous tentons ensuite de démontrer les limites d'un système interlingua. Nous montrons d'abord qu'une collaboration très étroite lors du développement des différentes grammaires nous permet généralement d'obtenir une même représentation sémantique dans les deux langues, ce qui rend le transfert plus efficace et offre la possibilité de construire une modélisation générale du domaine. Mais nous mettons ensuite en évidence, à l'aide d'exemples concrets, qu'une traduction de bonne qualité nécessite un système de transfert, qui seul nous permet de faire face à des variations structurelles et culturelles importantes.

2. Le projet de traduction automatique des bulletins d'avalanches

Ce projet a pour but l'implémentation d'un système de traduction automatique pour des bulletins d'avalanches émis par l'IFENA[1], une à plusieurs fois par semaine pendant l'hiver. Ecrits initialement en allemand, ces bulletins doivent ensuite être traduits dans les deux autres langues officielles suisses, le français et l'italien. La possibilité de l'ouverture d'un Institut semblable en région francophone, qui rédigerait les bulletins en français, nous a décidés à adopter une approche bidirectionnelle. Pour l'instant, le travail se concentre sur le français et l'allemand. Nous avons déjà montré que les bulletins d'avalanches constituent une application idéale pour la traduction automatique (BOUILLON et al., 1991 et 1992a, BOESEFELDT et al., 1991). Ils traitent en effet du domaine limité des avalanches en Suisse et utilisent un sous-langage bien défini (SALM, 1982), quoique relativement complexe. Il est donc possible de limiter le traitement automatique à ce sous-langage et de bénéficier de tous les avantages qu'il offre. Nous pouvons notamment éviter de modifier le style des bulletins, en traitant tous les phénomènes du corpus et exclure toute post-édition humaine.

3. La mise en oeuvre des grammaires et le choix de la représentation sémantique

3.1. Le logiciel

Le système de traduction des avalanches est mis en oeuvre avec ELU, un environnement linguistique d'unification, développé à l'ISSCO pour interpréter des grammaires écrites dans un langage particulier. Il comprend quatre modules : un module lexical, un analyseur, un générateur et une composante de transfert (ESTIVAL, 1990). L'analyseur permet d'obtenir pour une phrase les représentations en structures d'attributs permises par la grammaire. Le module de transfert établit une relation binaire entre deux structures de traits pour permettre le passage d'une langue à l'autre (ESTIVAL et al., 1990b). Enfin, le générateur part d'une représentation en structures de traits et recherche dans la grammaire toutes les phrases qui peuvent être reliées à la structure initiale. Comme chacun de ces modules utilise la technique d'unification, ces trois étapes sont réversibles (RUSSELL et al., 1990 et RUSSELL et al., 1991).

[1] Ce projet est partiellement subventionné par l'IFENA (l'Institut Fédéral pour l'Etude de la Neige et des Avalanches).

3.2. Les grammaires

Depuis 1990, le travail a porté sur la construction de deux grammaires allemande et française, pour le traitement de tous les phénomèmes syntaxiques rencontrés dans le corpus des avalanches. Cette étape est loin d'être triviale. Les bulletins présentent en effet un grand nombre de problèmes syntaxiques, bien connus, mais peu implémentés, comme celui de la coordination, des subordonnées, des mots composés allemands, de la topicalisation, des temps, etc. Bien que les stratégies adoptées pour l'écriture des grammaires allemande et française diffèrent considérablement, nous avons essayé, lorsque c'est possible, d'obtenir la même représentation sémantique dans les deux langues. Une simplification du transfert, moins bien expérimenté jusqu'ici, permet en effet de limiter le nombre de règles de transfert et de gagner en efficacité.

3.3. Une représentation sémantique orientée interlingua

La liaison entre les expressions linguistiques allemandes et françaises qui traitent d'un même contenu s'effectue donc au niveau de la représentation sémantique. Elle permet de faire abstraction des particularités syntaxiques de chaque langue et de ne représenter que les informations nécessaires à la traduction. Comme le domaine traité est bien modélisé, il est possible de déterminer à l'avance les différentes composantes sémantiques des phrases du corpus des avalanches, comme la description du danger, le type de neige, le lieu, le climat, le temps, etc. Avec le logiciel ELU qui utilise l'unification, la représentation sémantique se présente sous la forme de structures d'attributs (SHIEBER, 1986). Dans notre projet, les traits PRED, ARGS, MOD, MORPH et POSITIF permettent d'encoder respectivement le prédicat logique de la phrase, ses arguments, les différents modifieurs de la phrase, les indications morphologiques nécessaires à la traduction, comme le temps, la voix et le caractère positif ou négatif de la phrase.

Par exemple, la phrase *dans les Alpes subsiste un danger d'avalanches* recevra la structure sémantique suivante :

```
args = [ <81> compl : compl = []
                      detype = non
                      pred = avalanches
           pred = danger
           detype = indefinite
           mod : alt : pred = []
                 expo : pred = [] ]
mod : alt : pred = []
      loc : pred = Alpes
            detype = definite
            top = yes
      temps : pred = []
      climat : pred = []
morph : temps = present
        voix = actif
positif = yes
pred = subsister
```

Cette structure signifie que la phrase est positive, au présent actif et indique que le prédicat logique est le verbe intransitif *subsister*. Celui-ci a pour argument le sujet indéterminé *un danger d'avalanches*, encodé dans une liste. Un seul modifieur est réalisé, le complément de lieu *dans les Alpes*. Le trait <* top> = yes signale qu'il est topicalisé. Les listes vides ([]) indiquent que la phrase ne contient pas d'autres modifieurs.

Dans ce cas-ci, l'équivalent allemand *in den Alpen besteht eine Lawinengefahr* recevra la même structure sémantique. Seules les valeurs des traits, les mots français, sont remplacées par d'autres valeurs : les mots allemands.

Pour obtenir une même représentation dans les deux langues, nous avons dû complexifier les grammaires et le lexique. Dans le cas de l'exemple cité plus haut, les lexiques et les grammaires doivent contrôler les prépositions de lieu, qui ne font pas partie de la représentation sémantique. Tous les systèmes qui essayent d'établir une correspondance entre les prépositions se heurtent en effet à des problèmes importants et sont obligés de définir des algorithmes très complexes, qui, dans notre cas, ralentiraient considérablement le système (JAPKOWICS, 1991). De même, les mots composés allemands sont décomposés dans le lexique, de manière à obtenir la même représentation qu'en français. Cette technique nous permet un traitement général et unifié du problème des noms composés (BOUILLON et al., 1992b). Enfin, le mode et le genre, qui varient en fonction des langues ont été exclus de la représentation sémantique.

Parfois, l'obtention d'une même représentation est loin d'être aisée. Prenons l'exemple de la coordination des syntagmes nominaux. Tandis que le français a tendance à répéter l'article devant les différents noms qui composent le syntagme nominal coordonné et permet très rarement l'élision de l'article à l'intérieur du syntagme coordonné, l'allemand peut faire porter l'article sur tout le groupe coordonné dans différents cas :

- *Schneeverfrachtungen führten zu einer leichten Setzung und Verfestigung der Schneedecke*
- *des accumulations de neige ont causé une consolidation et un tassement légers de la couverture de neige.*

Dans une optique monolingue, nous traduirions cette différence syntaxique au niveau de la représentation sémantique du groupe coordonné. En allemand, le trait

detype qui encode le type de l'article se trouverait en dehors de la liste des arguments (1), ce qui permettrait de distinguer cette construction de celle qui implique une répétition de l'article, alors qu'en français, ce dernier serait toujours répété à côté de chaque élément de la liste (2) :

```
(1)  args : [ <1> pred = setzung
              <2> pred = verfestigung ]
     detype = indefinite
     mod : pred = leicht
     pred = und

(2)  args = [ <1> detype = indefinite
                   pred = tassement
              <2> detype = indefinite
                   pred = consolidation ]
     mod : pred = léger
     pred = et
```

Dans une optique multilingue, cette différence dans la représentation constitue un problème au niveau du transfert. Elle nous oblige à écrire autant de règles de transfert qu'il peut y avoir d'éléments à l'intérieur de la liste, ce qui est peu général et restrictif. De plus, une étude du corpus montre que les deux types de construction ne sont jamais utilisés dans le même contexte et qu'il est possible de définir les conditions de rejet et d'acception de chacune de ces structures dans le cadre du sous-langage des bulletins d'avalanches. Dans ce cas-ci, nous avons donc préféré simplifier le transfert et complexifier la grammaire allemande, pour obtenir la même représentation qu'en français, avec répétition du trait <* detype> = indefinite à côté de chaque élément de la structure de liste allemande, comme suit:

```
args : [ <1> detype = indefinite
              pred = setzung
         <2> detype = indefinite
              pred = verfestigung ]
mod : pred = leicht
pred = und
```

3.4. Les limites de la représentation interlingua

La réalisation d'une même représentation s'avère cependant beaucoup plus difficile dans un certain nombre de cas. Deux expressions de langues différentes n'expriment en effet pas nécessairement de la même manière un fait identique. Tout d'abord, une réalité peut être plus ou moins importante en fonction du contexte culturel et social dans lequel la langue évolue. D'autre part, une langue peut offrir plus de possibilités syntaxiques ou sémantiques qu'une autre.

Dans le corpus des avalanches, de telles différences sont évidentes et nous allons le montrer à l'aide de quelques exemples.

Prenons d'abord en considération une divergence temporelle. Alors que les bulletins allemands utilisent indifféremment l'imparfait et le passé composé, pour désigner un passé composé français, les bulletins français ne contiennent aucun imparfait, temps réservé pour des faits en train de se dérouler dans la durée, exclus de la réalité présente :

- *am Alpensüdhang fielen 80 cm Schnee*
  -> sur le versant sud des Alpes sont tombés 80 cm de neige
- *am Alpensüdhang sind 80 cm Schnee gefallen*
  -> sur le versant sud des Alpes sont tombés 80 cm de neige.

Cette habitude en allemand peut s'expliquer de deux manières : d'une part, l'imparfait est plus aisé à former et permet un accès plus rapide à l'information lexicale. D'autre part, les bulletins sont écrits par des locuteurs du suisse allemand, qui peuvent avoir tendance à beaucoup utiliser l'imparfait en allemand, inusité dans leur dialecte. Pour traiter cette différence, diverses solutions étaient envisageables. Nous pouvions interdire l'utilisation de l'imparfait en allemand, ce qui est peu élégant et témoigne d'un manque de souplesse. Nous pouvions aussi éviter une telle restriction et définir deux règles de transfert qui établissent respectivement une correspondance entre le passé composé allemand et le passé composé français et entre l'imparfait allemand et le passé composé français. Dans ce cas, nous devions aussi bloquer la réversibilité de la seconde règle, pour empêcher la génération de deux solutions en allemand. Dans la syntaxe ELU, ces règles présentent la forme suivante :

```
:T:  tempo1
:L1: <* morph temps> = passe_comp
:L2: <* morph temps> = passe_comp
:X: -

:T:  tempo2
:L1: <* morph temps> = imparfait
:L2: <* morph temps> = passe_comp
     <* reversibilite> = non
:X: -
```

La première, **tempo1**, établit une correspondance entre les passés composés en allemand (L1) et en français (L2). La seconde, **tempo2**, transforme l'imparfait allemand en un passé composé français. Ces règles s'appliquent si la représentation de la langue source est subsumée par la structure de traits décrite dans L1 et si la représentation pour la langue cible unifie avec la structure de traits définie dans L2 (ESTIVAL et al.(1990b) et RUSSELL et al.(1991)).
Le trait <* reversibilite> = no, qui ne sera jamais subsumé par une structure de traits française, bloque donc la réversibilité de cette règle (RUSSELL et al., 1991). Ainsi, tous les passés composés et les imparfaits allemands se traduiront par des passés composés français et le passé composé français ne se traduira que par le passé composé allemand, ce qui semble à nos yeux la meilleure traduction.

Un problème similaire se pose, quand nous voulons traduire le participe présent allemand. Alors qu'en allemand, le participe présent est couramment utilisé, le français a tendance à le remplacer par une relative.

Par exemple, la phrase suivante :

- *die anhaltenden Niederschläge und die Setzung der Schneedecke führten zu einer Abnahme der Lawinengefahr*

contient le participe *anhaltenden* qui se traduira de préférence en français par la relative *qui continuent* :

- *les précipitations qui continuent et le tassement de la couverture de neige ont causé une diminution du danger d'avalanches.*

Comme les relatives existent aussi en allemand, il est peu souhaitable d'obtenir la même structure en français et en allemand, ce qui provoquerait une surgénération. Nous avons donc choisi de créer une règle de transfert, qui établit une correspondance entre la structure allemande :

```
pred = Niederschläge
mod : pred = anhalten
rel = []
```

et la structure française correspondante, où le signe #12 indique une structure réentrante :

```
pred = #12 précipitation
mod = []
rel : args : [ pred = #12 ]
     pred = continuer
```

Comme ces deux structures sont assez différentes, la règle est relativement complexe : elle stipule que le prédicat du modifieur en allemand Z1 correspond au prédicat de la relative Z2. Cette relative a pour argument une liste [R], dont le prédicat X2 est semblable au prédicat de la phrase nominale (réentrance) et correspond au nom auquel se rapporte le participe allemand X1.

```
:T: part_rel
:L1: <* pred> = X1
     <* rel> = []
     <* mod pred> = Z1
:L2: <* pred> = X2
     <* mod> = []
     <* rel> = W
     <W args> = [R]
     <R pred> = X2
     <W pred> = Z2
     <* reversible> = non
:X: X1 = X2
    Z1 = Z2
```

La coordination aussi exige un traitement semblable. Alors que l'allemand utilise indifféremment la virgule ou la conjonction **und** pour coordonner deux adjectifs, le français ne permet que la conjonction **et** :

- *der feuchte, instabile Schnee hat zu einer ernsthaften Lawinensituation geführt*
- *der feuchte und instabile Schnee hat zu einer ernsthaften Lawinensituation geführt*
- *la neige instable et humide a causé une grave situation d'avalanches*
- **la neige instable, humide a causé une grave situation d'avalanches*

Pour permettre les deux constructions en allemand, il est donc indispensable d'établir une correspondance entre ces deux structures :

```
(1)  mod : [<1> args : [ <2> pred = feucht
                         <3> pred = instabil ]
                pred = , ]
(2)  mod : [<1> args : [ <2> pred = humide
                         <3> pred = instable ]
                - pred = et ]
```

La règle de transfert suivante établit cette correspondance. Elle stipule que si le trait MOD en allemand a pour valeur une liste dont le prédicat est la virgule, nous obtiendrons une liste similaire en français avec, pour prédicat, la conjonction **et**.

```
:T: virget
:L1: <* mod> = [A]
     <A pred> = ','
     <A args> = X
:L2: <* mod> = [B]
     <B pred> = et
     <B args> = Y
     <B bidirectionnel> = non
:X: X = Y
```

Cette règle n'est pas bidirectionnelle parce que nous ne voulons pas que tous les **et** français se traduisent par des virgules en allemand.

Enfin, pour ne citer qu'un dernier exemple, un grand nombre d'adjectifs en allemand se traduisent de préférence par des noms en français :

- *die östlichen Alpen* -> la partie est des Alpes
- *die mittleren Alpen* -> le centre des Alpes

L'utilisation d'adjectifs serait aussi possible dans les traductions françaises, mais elle n'est pas conforme aux habitudes langagières et doit de ce fait être évitée dans le cadre de bulletins d'avalanches.

Nous avons donc préféré définir une règle de transfert qui transformera la structure allemande suivante :

```
pred = Alpen
detype = definite
mod : pred = östlich
```

en une structure qui permettra la génération des traductions proposées ci-dessus :

pred = partie
detype = definite
mod : pred = est
compl : pred = Alpes
detype = definite

Notons que le syntagme *les Alpes centrales* ne sera pas exclu pour autant par la grammaire française. L'analyseur produira une structure semblable à celle de l'allemand et ce syntagme sera lui aussi traduit en allemand par *die mittleren Alpen*.

4. Conclusion

Dans cet article, nous avons mis en évidence un certain nombre de problèmes intéressants de traduction, auxquels nous sommes confrontés pour le traitement automatique des bulletins d'avalanches de la Suisse et nous avons montré comment le logiciel ELU permet de les résoudre. Même si une collaboration étroite lors de l'écriture des grammaires permet d'obtenir une représentation sémantique cohérente, qui modélise le domaine des avalanches en Suisse, nous maintenons que, pour faire face à des variations structurelles ou des habitudes langagières, seul un système basé sur le transfert permet d'obtenir une traduction de qualité.

5. Bibliographie

Boesefeldt (K.) et Bouillon (P.) (1991).- Le rôle de la représentation sémantique dans un système de traduction multilingue.- in: Working Paper 58, ISSCO, 1991, Genève.

Bouillon (P.) et Boesefeldt (K.) (1991).- Applying an Experimental MT System to a Realistic Problem.- in: Proceedings of Machine Translation Summit III, Washington, July 1991, pp. 45-49.

Bouillon (P.) et Boesefeldt (K.) (1992a).- La Traduction automatique des Bulletins Avalanches.- à paraître in: Colloque International sur L'Environnement Traductionnel, Mons, 1992.

Bouillon (P.), Boesefeldt (K.) et Russell (G.) (1992b).- Compound Nouns in a Unification-Based MT System.- in: Proceedings of 3rd Conference on Applied Natural Language Processing, Trento, March-April 1992, pp. 209-215.

Estival (D.) (1990a).- Elu User Manual.- in: Technical Note 1, ISSCO, Genève, 1990.

Estival (D.) (1990b), Ballim (A.), Russell (G.) and Warwick (S.).- A Syntax and Semantics for Feature-Structure Transfer.- in: The Third International Conference on Theorical and Methodological Issues in Machine Translation of Natural Language, 1990.

Japkowicz (N.) et Wiebe (M.).- A System for Translating Locative Prepositions from English to French.- in: 29th Annual meeting of The Association for Computational Linguistics, Berkeley, 1991

Russell (G.), Ballim (A.), Estival (D.) et Warwick (S.) (1991).- A Language for the Statement of Binary Relation over Feature Structures.- in: Proceedings of European Association for Computational Linguistics, 1991.

Russell (G.), Carroll (J.) et Warwick (S.) (1990).- Asymmetry in Parsing and Generating with Unification Grammars: Case Studies from ELU.- in: Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, June 1990, pp. 205-211.

Salm (B.) (1982).- Lawinenkunde für den Praktiker.- Bern: Verlag des SAC, 1982.

Shieber (S.M.) (1986).- An Introduction to Unification Based Grammar.- in: CSLI Lecture Note No. 4, 1986.