RECITAL 2004, Fès, 21 avril 2004

Génération sémantico-syntaxique pour la traduction automatique basée sur une architecture interlingue

Mehand Iheddadene
Thèse à France Telecom R&D - DMI/GRI
2, avenue Pierre Marzin - 22307 Lannion
mehand.iheddadene@rd.francetelecom.com

Résumé - Abstract

Dans cet article, nous présentons un processus de génération sémantico-syntaxique conçu et mis en œuvre dans la réalisation d'un prototype de traduction automatique basée sur le modèle à structure intermédiaire (ou structure pivot). Dans une première partie de l'article, nous présentons l'organisation des ressources lexicales et sémantiques multilingues, ainsi que les mécanismes permettant d'exploiter ces ressources pour produire une représentation conceptuelle du sens de la phrase source. Dans une seconde partie, nous présentons la première phase de génération à partir d'une structure pivot (génération Sémantico-Syntaxique) permettant la construction d'une structure syntaxique profonde de la phrase cible à produire. Les autres phases de génération ne seront pas abordées dans cet article.

This paper aims to present a method for the semantic-syntactic generation that has been proposed to build a machine translation prototype based on the interlingua translation model. The first part of the paper introduces the multilingual lexical and semantical resources, and the way they are used to build conceptual representations of the meaning of source sentences. The following part of the paper presents the first step in generating the target text from the conceptual representation (semantic-syntactic generation) which results on a deep syntactic structure of the targeted text. The other steps of the generation process will not be discussed in this paper.

Mots-clefs – Keywords

Traduction Automatique, Architecture interlingue, sémantique lexicale, Génération sémantico-syntaxique, structure syntaxique profonde
Machine Translation, Interlingual architecture, lexical semantics, semantic-syntactic generation, deep syntactic structure

1 Introduction

La traduction automatique basée sur une structure pivot (ou interlingue), est explorée dans un certain nombre de systèmes et prototypes : ULTRA (Farwell & Wilks, 1991), DLT de BSO Research (Witkam, 1988), projet UNL (Sérasset & Boitet, 1999), NESPOLE (Metze *et al.*, 2002), C-STAR, etc. Une telle structure pivot peut se matérialiser dans des conceptions différentes allant d'une langue arbitraire telle que l'Espéranto (utilisée par le système DLT) à une représentation plus au moins abstraite dans une langue donnée telle qu'un arbre syntaxique ou sémantique en anglais (Boitet, 2000). Elle peut aussi être représentée par des ontologies (système KBMT (Nirenburg *et al.*, 1986) et le projet KANT (Mitamura *et al.*, 1991)). Dans notre approche, nous nous situons dans le cadre de l'application de la théorie sens-texte (TST) ((Mel'čuk, 1997), (Polguère, 1998)) à la traduction automatique. La TST fournit justement les moyens de modélisation d'un langage pivot reposant sur l'hypothèse que le sens d'une phrase source, après une analyse syntactico-sémantique assez profonde, peut être représenté au moyen d'un langage non spécifique entièrement indépendant des langues.

2 Modélisation des informations sémantiques

La représentation intermédiaire proposée dépend essentiellement de la représentation des unités lexicales qui composent la phrase source. Nous considérons que la description de l'unité lexicale et sa formalisation doivent être réalisées de telle sorte que le système automatique puisse prédire, à la fois son comportement au sein de l'énoncé et sa relation avec les unités co-occurrentes. Il est donc primordial de commencer par décrire la manière de formaliser les informations sémantiques, voire encyclopédique à inclure dans le lexique. Cette modélisation est formalisée à l'aide d'un thésaurus constitué d'un ensemble de thèmes multilingue. Un thème est un ensemble d'entrées lexicales (ou lexicalisations) abordant le même sujet. Ainsi, nous regroupons, par exemple, les unités lexicales *chien* et *dog* dans le même thème, *pharaon, empereur* et *roi* dans un autre thème et *lire, lecture* et *lecteur* dans un autre, etc. Le découpage en thèmes constitue une partition de l'ensemble des sens des entrées lexicales des langues traitées. En conséquence, si une entrée lexicale apparaît dans plusieurs thèmes, ces apparitions seront considérées comme plusieurs sens de cette entrée. Les éléments constituant les thèmes sont appelés synsets (inspirés de WordNet) ou blocs sémantiques. L'utilisation d'un tel thésaurus dans des applications de traitement automatique des langues nécessite la mise en œuvre d'un modèle sémantique.

2.1 Modèle Sémantique

Ce modèle (Vinesse, 2000) s'inspire de l'organisation des vocables dans le Dictionnaire Explicatif et Combinatoire ou **DEC** ((Mel'čuk *et al.*, 1995)). Il permet de représenter le potentiel paraphrastique du lexique en permettant à la fois : (1) d'obtenir, lors de l'analyse, des représentations sémantiques identiques pour des énoncés équivalents et (2) de proposer plusieurs paraphrases pour une même représentation sémantique lors de la génération. Le modèle sémantique proposé par France Telecom R&D repose sur le concept de **tribus**[1]. Une tribu

[1]Le concept des tribus a été proposé par J.Vinesse, responsable du groupe de TALN à France Telecom R&D.

est composée d'un ensemble de blocs sémantique. Dans un bloc sémantique, on trouve seulement les lexicalisations partageant, en plus d'une interprétation sémantique commune, un même comportement syntaxique, i.e. ils doivent avoir des régimes syntaxiques identiques à savoir : le même nombre de positions syntaxiques et une équivalence de ces positions une à une et dans le même ordre. Ainsi, les expressions "F.musique" et "S.musica", (voir figure 1), appartiennent au même bloc sémantique[2], car elle partagent, outre un sens commun, un même comportement syntaxique (toutes les deux sont des *noms* réalisant la même position syntaxique). Par contre "F.musicien" et "F.musique", bien qu'étant toutes deux des noms, elles ne sont pas dans le même bloc sémantique car leurs régimes syntaxiques diffèrent. Formellement, chaque tribu est constituée d'un ensemble de prédicats permettant de décrie les différentes distributions d'arguments possibles auxquelles seront associées les lexicalisations qui figurent dans la tribu en question. Ainsi la distribution argumentale du prédicat INSTRUMENTAL_OPERATION.musique possède trois arguments : *situation, agent* et *patient*. "F.musicien" et "E.musician" sont des lexicalisations de l'argument *agent*, "F.musique" lexicalise le *patient* et "F.musiquer" lexicalise à la fois les deux arguments *situation et agent*. Pour le prédicat MUSICAL_INSTRUMENT_OPERATION.piano, il n'existe pas (en français) de lexicalisation pour l'action de *"jouer du piano"*, mais il y en a une pour l'agent de cette action ("pianiste"). Nous distinguons deux types de prédicats : (1) un prédicat de base (INSTRUMENTAL_OPERATION.musique dans l'exemple) et (2) d'autres prédicats qui sont déduits à partir du premier en utilisant ce que nous appelons les Fonctions de Lexicalisation (FdL). Ces Fonc-

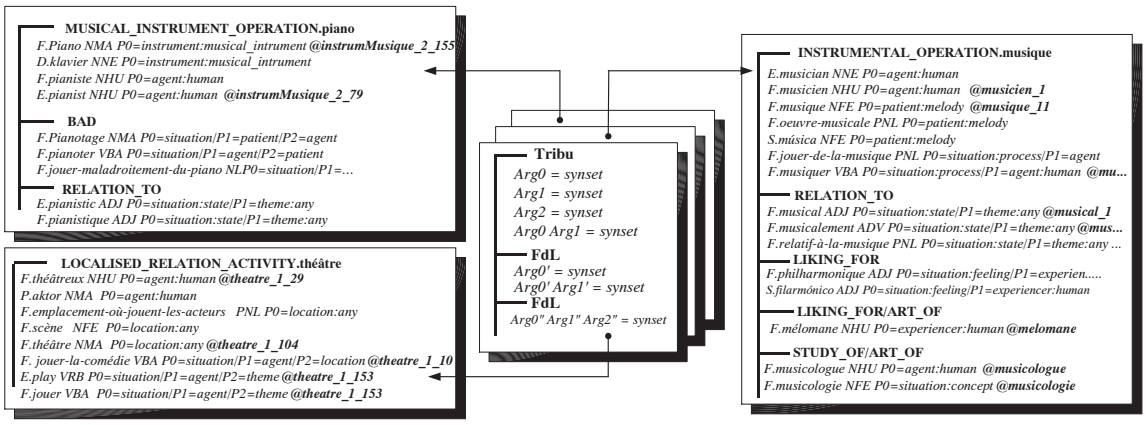

Figure 1: Extraits de tribus

tions de Lexicalisation permettent de passer d'un sens à un autre au sein d'une même tribu en regroupant les lexicalisations ayant un ensemble de caractéristiques sémantiques communes. Ces lexicalisations donnent, à travers les FdLs, les différents points de vue sur le prédicat de base. Ainsi, par exemple, nous trouvons dans la même tribu les deux lexicalisations "F.musicien" et "F.mélomane" reliées par la Fonction de Lexicalisation LIKING_FOR/ART_OF qui signifie <aimer l'art de> (*mélomane* étant la personne passionnée de l'art que pratique le *musicien*). Une FdL n'est pas propre à une tribu donnée, au contraire une Fonction de Lexicalisation n'est considérée comme telle que si elle est jugée suffisamment fréquente et reproduit cette même relation au sein d'autres tribus. Ces FdLs permettent, à la fois, de donner le sens de la lexicalisation et d'expliciter ses relations avec les autres lexicalisations de la même tribu.

[2]La délimitation des blocs sémantiques à l'intérieur de la tribu n'est pas explicitée par une marque spécifique. Elle apparaît seulement en comparant les structures argumentales des différentes lexicalisations.

2.2 Structure Pivot Interlingue

Dans notre approche, nous considérons qu'il est nécessaire d'introduire des mécanismes permettant à la fois d'expliciter la structure phrastique de l'énoncé à traduire et d'associer à chaque élément de cet énoncé une description indépendante des langues. Le moyen choisi pour réaliser cet objectif est l'utilisation d'une structure interlingue modélisée par un graphe sémantique. Elle permet d'identifier les éléments de sens individuels constituant la phrase source, leur structure argumentale (prédicat à un, deux,... arguments) et d'établir les connexions predicat-argument unissant ces différents éléments de sens. Le graphe sémantique de la figure 2 est la structure interlingue des paraphrases ayant pour invariant de sens commun celui véhiculé par *"the pianists play music with pianos"*. Les graphes sémantiques associés à une phrase donnée sont obtenus

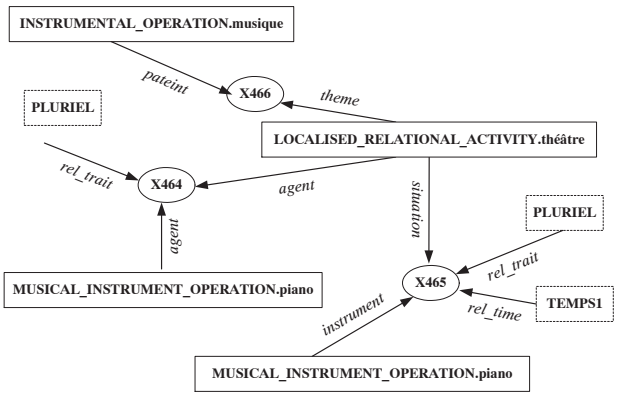

Figure 2: Un graphe sémantique

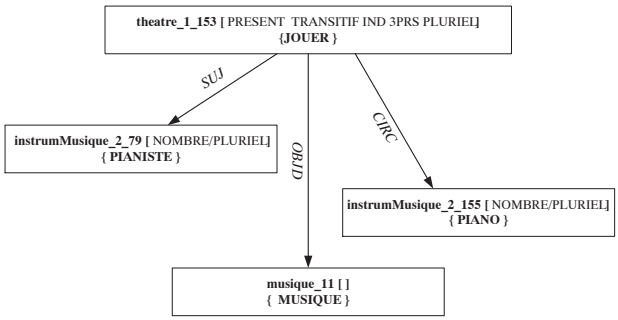

Figure 3: Structure Syntaxique Profonde

à partir de l'analyse en dépendance de cette phrase et des ressources citées plus haut. En effet, les éléments déclenchant les règles de construction de la structure interlingue sont les fonctions syntaxiques présentes dans l'arbre de dépendance. Chaque couple d'unités lexicales pleines liées par une fonction syntaxique fait émerger dans la structure interlingue, par le biais de leur unité sémantique, des prédicats. Ceux-ci seront reliés entre eux par des actants sémantiques calculés en utilisant leur schéma prédicatif. D'autres règles sont utilisées pour compléter les graphes sémantiques par des informations aspectuelles et communicatives (topic, focus, etc.) en inférant, à partir de certains traits (syntaxiques et sémantiques) associés aux unités lexicales, des prédicats génériques (PLURIEL, TEMPS1, etc.).

3 Génération Sémantico-Syntaxique

Dans ce qui suit, nous allons décrire uniquement[3] le premier module de génération (module de génération Sémantico-Syntaxique). Il permet de générer une structure syntaxique profonde à partir d'un graphe sémantique. Cette structure (figure 3) se présente sous la forme d'un arbre de dépendance dont les arcs sont étiquetés par des fonctions syntaxiques reliant des lexèmes pleins rattachés aux mots dans le lexique monolingue de la langue cible. Chaque lexème est accompagné par une suite de grammèmes. Ce module prend en charge les éléments suivants :

[3]Les autres modules de l'architecture stratifiée globale du générateur ne seront pas décrits dans cet article (génération des structures syntaxiques de surface, ordonnancement syntaxique des mots, génération morpho-lexicale des formes fléchies des mots linéarisés, etc.).

1- L'étape d'hiérarchisation a pour objectif de hiérarchiser le graphe sémantique par un parcours permettant d'obtenir une structure d'arbre. Si aucun choix n'est fait sur le nœud de début de parcours, des risques d'explosion combinatoire peuvent se produire sur le nombre d'arbres produits. Il est donc indispensable de calculer le nœud qui sera la tête de la phrase à générer[4]. Dans la version actuelle, nous considérons que le nœud tête de la phrase cible est le même que celui de la phrase source. Une méthode de calcul sera mise en œuvre en utilisant les notions de focus, topic et perspective. L'algorithme de parcours du graphe sémantique transite une seule fois par chaque actant sémantique (i.e. arc) et constitue, à chaque transition, une branche de la structure profonde. Lorsque l'algorithme traverse le même nœud plus d'une fois, un lien de co-référence est alors créé dans la structure syntaxique profonde.

2- Le module de lexémisation a pour rôle de remplacer les prédicats interlingues par des lexèmes de la langue cible qui seront ensuite lexicalisés, dans le module de génération syntaxique, en unités lexicales appropriées. Le choix des lexèmes et des relations syntaxiques est dicté par les informations contenues dans le modèle sémantique, le lexique monolingue et la grammaire de la langue cible. Comme cité plus haut, les unités lexicales appartenant à la même famille sont regroupées au sein de la même tribu. Ainsi, à partir d'une tribu et d'un schéma prédicatif donné, nous pouvons déduire toutes les lexicalisations paraphrastiques réalisant chacun des arguments de ce schéma prédicatif. En effet, dans un premier lieu, nous récupérons, à partir du modèle sémantique de la langue cible, tous les lexèmes appartenant au même prédicat. Ensuite, un tri est effectué pour garder seulement ceux qui ont la même distribution argumentale que celle identifiée dans la structure interlingue.

3- Les relations syntaxiques sont déduites, d'une part, des tableaux de régimes associés pour chaque unité lexicale stockés dans le lexique sous forme d'un trait appelé FONCPOS et d'autre part, du schéma prédicatif associé à leurs lexèmes. Le lien est fait grâce aux positions profondes P0, P1, etc. La fonction syntaxique étiquetant (dans la structure profonde) l'arc reliant les deux lexèmes réalisant le même rôle, est calculée en utilisant une règle grammaticale propre à la langue cible. Cette règle combine les deux relations syntaxiques associées à ces deux lexèmes.

4- Le module utilisant les règles sémantico-syntaxiques à pour objet d'inférer, à partir des prédicats génériques, des grammèmes qui seront transmis aux niveaux postérieurs (générations syntaxique, lexicale, morphologique,...). Ceci permet de contrôler le calcul de certaines constructions grammaticales et de certains indicateurs morphologiques (temps grammatical, nombre, genre, défini/indéfini,...) ou encore de faire émerger les lexèmes dits vides (articles, verbes auxiliaires,...). Ce module utilise un ensemble restreints de règles sémantico-syntaxiques dont le modèle est représenté sur la figure 4. Celui-ci met en correspondance un sous graphe qui contient le rôle déclencheur de la règle et le prédicat avec le trait associé (Predicat_x(trait)).

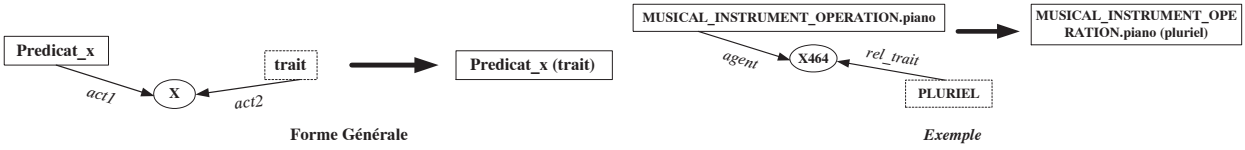

Figure 4: Règle sémantico-syntaxique

[4]Nœud dominant selon (Polguère, 1990) qui propose une méthode de calcul de ce nœud basée sur l'opposition thème/rhème de la structure communicative.

4 Conclusion

Les ressources nécessaires à un logiciel de traduction automatique comprennent toujours un lexique et une grammaire. La construction de celles-ci est un processus incrémental souvent long, cependant, nous avons atteint une couverture suffisamment importante pour procéder à la première série de tests. Au 06 Mars 2004, le nombre de tribus construites était de 4940 tribus contenant 34372 unités lexicales multilingues dont 12571 en français et 8470 en anglais. Certains domaines (culture, agriculture, alimentation, etc.) sont majoritairement couverts. Les premiers résultats sont encourageants et plusieurs indications pour résoudre les problèmes détectés sont en phase de mise en œuvre en exploitant le modèle sémantique présenté dans cet article et particulièrement les Fonctions de Lexicalisation.

Références

BOITET C. (2000). Traduction assistée par ordinateur. In J. M. PIERREL, Ed., *Ingénierie des langues*, chapter 12, p. 271–291. HERMES Science Europe.

FARWELL D. & WILKS Y. (1991). Ultra: a multi-lingual machine translator. In *Proceedings of the MT Summit III*, Washington, USA.

MEL'ČUK I. (1997). Vers une linguistique Sens-Texte, leçon inaugurale. *Paris, Collège de France*.

MEL'ČUK I., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*, volume 1. Editions Duculot, AUPELF-UREF.

METZE F., MCDONOUGH J., SOLTAU H. & WAIBEL A. (2002). The nespole! speech-to-speech translation. In *Proceedings of the Second International Conference on Human Language Technology*.

MITAMURA T., NYBERG E. H. & CARBONELL J. G. (1991). An effecient interlingua translation system for multilingual document production. In *Proceedings of the MT SUMMIT III*, p. 55–61.

NIRENBURG S., RASKIN V. & TUCKER A. (1986). On knowledge-based machine translation. In *Proceedings of the COLING86*, p. 627–632, Columbia University New York, New York, USA.

POLGUÈRE A. (1990). *Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte*. PhD thesis, Université de Montréal, Faculté des études supérieures.

POLGUÈRE A. (1998). La Théorie Sens-Texte. *Dialangue, 8-9, Université du Québec à Chicoutimi*.

SÉRASSET G. & BOITET C. (1999). UNL-french deconversion as transfer & generation from an interlingua with possible quality enhancement through offline human interaction. In *Proceedings of the Machine Translation Summit VII*, Singapore.

VINESSE J. (2000). *THEMA, modèle de sémantique lexicale*. Rapport interne, France Telecom R&D, Lannion, France.

WITKAM T. (1988). DLT - an industrial R&D project for multilingual MT. In *Proceedings of the COLING88*, Budapest.