XML for a Unified Processing of Multilingual Corpora and Corresponding Translation Rules

Pascaline Merten

HEB (Haute école de Bruxelles)
ISTI (Institut supérieur de traducteurs et interprètes)
34, rue Hazard – B-1180 Bruxelles
pmerten@heb.be

Abstract

Our aim is to determine the rules which govern the order of the French modifiers. It also has a comparative dimension where French is the target language. Consequently, we formalized bilingual corpora, as well as translation rules and lexicons. To achieve this aim, we exploited the XML language as well as other languages of the same family: XPath and XLink. The procedures were written in XSLT. We also took benefit from the lessons of the Text Encoding Initiative (TEI) and HyTime, in particular with regard to the notion of hyperlink.

1 Introduction

Our framework is the following:

- we adopted the point of view of a human or machine translation (MT) system, with French as target language;
- within this framework, we seek to establish the rules that govern the order of the modifiers;
- we adopted also a comparative point of view, since we considered what occurred in various source languages, and various specialized domains;
- we also wanted to test those rules, by implementing them.

This study led us to to examine the various means of standardizing and formalizing both the bilingual corpora and the translation rules.
We built a small implementation of these rules in an XSLT stylesheet, on a corpus of examples. The stylesheet exploits transfer and generation data contained in the corresponding lexicons.
After a brief overview of the linguistic problem, we will focus on the XML techniques.

2 The Order of Modifiers

The concept of modifier is a functional concept taken from Wilmet (2003). It widens the notion of adjective or epithet, as it includes the various morpho-syntactic categories which can modify the head noun:

- name: *fauteuil crapaud*[1]; *mother language, functional groups interconversions*
- relative clause: *les enfants qui sont sages; the man I saw yesterday*
- participles: *the corresponding rule/la règle correspondante; a distributed environment/un environnement distribué; het aangepaste formaat*/*le format adapté*
- prepositional phrase (PP) and genitives: *la porte de derrière; the girl with the red hair; my mother's car*
- adverb: *le toujours président*
- proper name (PN): *the Eiffel tower*.

This concept of modifier is particularly interesting in comparative linguistics, as a morpho-syntactic category is often tranlated by another category: *university degree* → *diplôme universitaire; machine translation* → *traduction automatique; the laughing man - > l'homme qui rit; une solution miracle* → *una soluzione miracolosa*, etc. In a TM, these expressions will be treated as transfer compounds, as these syntagms cannot be translated literally.
Many linguists working on Romance languages have analyzed the question of the *absolute* place of the adjective: that is its place before or after the name. Only a few have studied the *relative* order of the adjectives

[1] *Squat armchair*

and modifiers. More English linguists developped this point, as English, in comparison with other language, has an impressive ability to multiply premodifiers. But, despite the interest of these studies, they cannot be transposed as such in a language like French because

- morpho-syntactic categories differ, and the category has an impact on the absolute and relative orders (prepositional phrase and relative clauses cannot be anteposed; past participles are seldom anteposed…)
- word order rules are often based on semantic categories (*color*, *form*…): on one hand, these categories are too restricted to cover the whole variety of modifiers (an even of adjectives); on the other hand, semantic criteria are important, but are not the only ones.

It has also be argued that postmodifiers show a mirror order compared to the order of premodifiers. But this approach is also unsufficient: *individual N-protected aminoacid* → *acide aminé particulier N-protégé*; *serious insect infestation* → *prolifération inquiétante d'insectes*.

We will not present here the details of the linguistic word order rules, we will just give a few examples:

- the modifiers that express the nature of a concept precede the descriptive ones*: strenuous muscular activity during the race / activité musculaire intense durant la course*, *eendenborst met ananas* / *magret de canard à l'ananas*, *une robe de soirée/du soir décolletée/verte/longue*, *the central carbon atom / l'atome de carbone central*
- certain morphological forms are rejected at the end of the noun phrase: superlatives, recursive noun phrases, adjectives or participles governing a complement,: *soupe froide de crevettes de la Mer du Nord*; *la chaîne d'atomes de carbone la plus longue*
- The differences in absolute order has to be taken into account while translating: *grave errore del disco sul file* → *erreur disque sérieuse concernant le fichier*
- one of the hardest sequences to predict in French is the relative sequence of a PP and an adjective: *un chausson de danse rose* / ? *un chausson rose de danse*; *une compagnie de danse française / une compagnie française de danse*. Of course, these groups can be described as lexicalized units: *pomme de terre, collier de chien*… This is just a way to elude the problem, but it doesn't solve anything: what explains the ability of those expressions to form lexical units? In addition, many groups have an intermediate status, between the free group and the lexical unit. In that case the first rule should be applied.
- the order of the arguments of a nominalization is governed by other criteria: the general sequence is object-subject: *la production textile chinoise*, *la production de textiles par la Chine* but the morphology of the argument is also significant: *la production chinoise de textiles*.

3 XML Formalization

We studied the order of modifiers for translation purposes, so we built bilingual corpora, to perform comparative researches.

3.1 Bilingual Corpora

Several corpora were worked out by aligning monolingual files. The import-export format of the alignment softwares is the TMX format. TMX (Translation Memory eXchange), has been developed by the LISA organization[2]. This format describes a correspondence between a segment[3] in the source language and the corresponding segment in the target language:

```
<tu>
<tuv lang="EN-US">
        <seg>Hold both mouse buttons down for 2 seconds</seg></tuv>
<tuv lang="FR-FR">
        <seg>Appuyer sur les deux boutons de la souris pendant 2 secondes</seg></tuv>
</tu>
```

[2] "Localization Industry Standards Association", *see* http ://www.lisa.org/

[3] A segment is a unit used in translation technology, below the sentence level: a title, a cell, a portion of sentence delimited by a strong punctuation mark.

It is an implicit bond between the two segments. By associating an XSL stylesheet to the XML file, this file can be visualized as a table, which facilitates linguistic research and comparisons:

Hold both mouse buttons down for 2 seconds	Appuyer sur les deux boutons de la souris pendant 2 secondes
Do you accept all of the terms of the preceding License Agreement? If you choose to Decline; the application will close. You must Accept this agreement to continue.	Acceptez-vous les termes du contrat de licence précédent ? Si vous choisissez Refuser; l'application se fermera. Vous devez accepter ce contrat pour continuer.
Accept	Accepter
Decline	Refuser

Figure 1: Bilingual corpus in TMX format, associated to an XSLT stylesheet

3.2 Translation Process

In a machine translation context, the process is devided in 3 stages:

- the *analysis*, where a source segment is analyzed; the result is a tree structure
- the *transfer* where the source tree is transferred in a target tree: the words and the structures are translated and adapted to the target language; the output is an abstract tree structure in the target language
- and the *generation* where the final sequence is produced by applying rules specific to the target language.

Within the framework of this research, we considered only transfer and generation. The choice of the determiner, the elision and contraction of French determiners are typical generation problems : *erreur sur le disque / sur l'unité; coupable du délit de contrefaçon / s'assurer de la fiabilité*.

The order of modifiers is also a generation problem, since the criteria are not only semantic but also morphological and syntactic. And, if semantic characteristics may be considered as language independant, morphological and syntactic characteristics are specific to the target language.

The procedures call upon lexicons containing monolingual data necessary to the analysis and the generation, as well as bilingual data used to transfer the source tree and to translate transfer compounds.

3.3 The Representation of Grammatical Information

Formal grammars of the family of the unification grammars, widely used in natural language processing (NLP), highlighted the notion of grammatical information and the idea of information flow in the tree structure. The structures which we handle are dependency trees. For example, the number of the syntagm will have to percolate from the head to the modifiers :

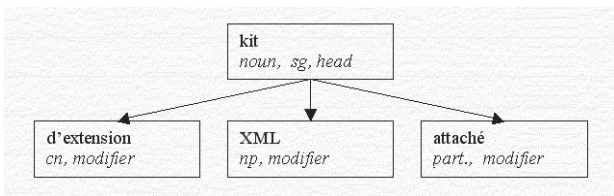

Figure 2: Dependency tree

The syntactic tree has an immediate correspondent in a XML tree structure. Functional units are represented by elements while other grammatical data are coded in attributes:

```
<sn id_sn="it_sn0" langue="en" id_tete="en_pack" nombre ="sg" quantif=" nil">
        <car cat="adj" source="en_attached" valeur1="descri" valeur2="etat"
        pos-absolue ="ante" pos-relative="3">attached</car >
        <car cat="np" source="en_XML" genre="m" valeur1="class" valeur2="nature "
        pos-absolue="ante " pos-relative="2">XML</car>
        <car cat="nom" source="en_expansion" genre="m" nombre ="sg" valeur1=" class"
        valeur2="nature " pos-absolue="ante" pos-relative="1">expansion</car>
        <tete cat="nom " source="en_pack">pack</tete>
</sn>
```

Note that a structure tree can be recursive.

3.4 Hyperlinks and the Representation of the Lexical Data

One of the main difficulties is the representation of correspondences between values located on different tree nodes. In order to achieve this, we used the XSLT variables, and XPath to describe tree paths.
We also benefit from the notion of hyperlink, in particular to code lexical data. This concept was described in very rich but sometimes different ways in the TEI project, in the HyTime and in the XLink languages.
Let us recall that a hyperlink connects anchors by means of pointers. The HTML hyperlink <a href=…> constitutes one of the anchors: it is a contextual and oriented link. It is described as a *simple* link. But the pointers of the hyperlink can also be conceived as attributes or subelements oh the hyperlink element. In that case, the hyperlink element doesn't constitute an anchor; it is out of context and can point to different targets : it is described as an *extended* link[4]. Both the simple and the extended links are explicit, while the correspondance markup used in the TMX format is implicit.

Here is a transfer lexical entry:

```
<tlex   id_tlex="n1"
        source="en_pack"
        cible="fr_kit" />
```

XLink recommends to employ subelements to represent the anchors. We could indeed, in a multilingual dictionary represent the same entry as follows:

```
<tlex   id_tlex="c4">
        <ancre langue=″en″ ancre_id="en_file_n" />
        <ancre langue=″fr″ ancre_id="fr_fichier" />
        <ancre langue=″nl″ ancre_id="nl_bestand" />
</tlex>
```

Whatever the technique, it is possible to express the semantics of the hyperlink and of its anchors, which is much richer than in HTML.

The identifiers of the transfer entry point to monolingual entries:

```
<lex    fr_id="fr_kit" cat="nom" genre="m" valeur="logiciel">
        <forme nombre="sg">kit</forme>
        <forme nombre="pl">kits</forme>
</lex>
```

Gender as well as semantic data are coded on the higher level (<lex> tag) as attributes, while the inflected forms are coded at a lower level (<form> tag), as element contents.

3.5 Representation of the Rules

There are three types of rules to be considered:

- transfer compounds
- standard transfer rules
- generation rules.

3.5.1 Transfer compounds

Here is an example of the transfer rule which ensures that the Italian phrase "errore del disco" will result in "erreur disque" in French:

```
<rtsf id_rtsf="rtsf1" id_it_tete="it_errore" id_it_car="it_disco">
        <cible><car source="fr_disque" cat="nom" nombre="sg" valeur1="class"
valeur2="nature"
        pos-absolue="post" pos-relative="1" /></cible>
</rtsf>
```

[4] Following the TEI, an extended link points to an external resource or a resource without identifier; XLink defines it as a hyperlink pointing to « an arbitrary number of resources » (DeRose, Maler et Orchard 2001 : $ 5.1.).

The rule needs one or two "prompters": those are coded by a direct, simple bond towards the identifier of the head and its modifier. The target is a local tree, that is to say a simple element or a local tree structure, which will be copied as it is in the output transfer tree.

3.5.2 Transfer and generation rules

Here is the transfer rule of an English modifying name into French. This English name will result in a PP in French, the tricky point is to identify its value: *a carbon atom* is translated by *un atome de ø carbone* whereas *the functional groups interconversions* results in *les interconversion entre les groupes caractéristiques*. The preposition can subcategorized by the head; if not, the preposition *de* will be used.

```
<transfert_caracterisants>
<car_source><car source="$src_car_id" cat="$src_car_cat" nombre="$src_car_nombre"
valeur1="$src_car_valeur1" valeur2="$src_car_valeur2" valeur3="$src_car_valeur3" pos-
absolue="$src_car_pos-absolue" pos-relative="$src_car_pos-relative" prep="$src_car_prep"
quantif="$src_car_quantif" /></car_source>
<car_cible><car source=
"$dico/dico/tlex[@en=$src_car_id]/@fr"cat="$dico/dico/lex[@en_id=$cible_car_id]/@cat">
<attributs      contrainte="$src_car_cat='nom'">
        <attribut       nom="genre" valeur="$dico/dico/lex[@en_id=$cible_car_id]/@genre"/>
        <attribut       nom="nombre" valeur="$src_car_nombre" />
        <attribut       nom="valeur1" valeur="$src_car_valeur1" />
        <attributs      contrainte="$src_car_valeur1='class'">
                <attribut       nom="cat" valeur="'cn'" />
                <attribut       nom="quantif" valeur="'nil'" /></attributs>
        <attributs      contrainte="$src_car_valeur1='specif' or $src_car_valeur1='argument'">
                <attribut       nom="cat" valeur="'cn'" />
                <attribut       nom="quantif" valeur="'ext'" /></attributs>
        <atributs>
        <attributs contrainte="$dico/dico/lex [@fr_id=$cible_tete_id]/sous-
        cat[@argument=$src_car_valeur2]">
                <attribut       nom="prep" valeur="$dico/dico/lex[@fr_id=$cible_tete_id]/
                sous-cat[@argument=$src_car_valeur2]/@prep" /></attributs>
                <attributs      contrainte="$dico/dico/lex[@fr_id=$cible_tete_id]/@sous-cat">
                        <attribut       nom="prep" valeur="$dico/dico/lex[@fr_id=$cible_tete_id]/
                        @sous-cat" /></attributs><attributs>
                        <attribut       nom="prep" valeur="'fr_de'" />
                </attributs></attributs></attributs>
...
</car></transfert_caracterisants>
```

Figure 3: A transfer rule

The transfer rule of the modifiers has two subelements: the source modifier and the target modifier. Various grammatical data are computed and stored as attribute values. The < attributes > elements apply when a certain constraint is met. In addition, these rules can be called recursively, to treat recursive phrases and sequences of modifiers.

The XSLT variables (marked $) inable to code shared values in different trees or sub-trees.

The "constraint" attribute of the "attributes" element implement the notion of grammatical constraint. If a certain constraint is met, one or several attributes will be generated and their values computed.

3.6 XSLT Implementation

The rules were implemented on different corpora in various language pairs and various subject fields. Here is the generation output for the example given above :

"attached XML expansion pack" →

```
<sn langue="fr" genre="m" nombre="sg" id_tete="fr_kit" quantif="nil">
  <tete cat="nom" source="fr_kit" valeur="logiciel" nombre="sg"
        genre="m">kit</tete>
  <car source="fr_attaché" cat="adj" valeur1="descri" valeur2="etat" pos-
        rel="post6">attaché</car>
  <car source="fr_XML" cat="np" valeur1="class" valeur2="nature" pos-
        rel="post4">XML</car>
  <car source="fr_extension" cat="cn" valeur1="class" valeur2="nature" pos-
        rel="post3" genre="" nombre="sg" prep="fr_de"
        quantif="nil">extension</car>
</sn>
```

→ "*kit d'extension XML attaché*"

4 Conclusion and Prospects

NLP must lay on a thorough linguistic analysis and has much to gain from a standard and declarative representation of data and rules. The contrastive approach and the implementation, beyond their practical interest force the analysis to be as rigorous as possible. The languages of the XML family seem particularly appropriate to the description of tree structures – such as they are employed in NLP – and to the description of grammatical information. In particular, rules describing linguistic and structural equivalences, make use of the capabilities of the hyperlinking and addressing mechanisms described in the XPath and XLink languages.
The representation of the lexicons should gain in generality as well as the rules in declarativity, but at least we hope we have demonstrated the possibility to code corpora and rules with markup languages of the XML family.

This research was completed thanks to the FNRS's financial support.

References

Bache, C. (1978). The Order of Premodifying Adjectives in Present-Day English. Odense University Press

Berglund, A., Boag, S., Chamberlin, D., Fernandez, M. F., Kay, M., Robie, J. and Siméon, J. (Eds) (2005). XML Path Language (XPath) 2.0. W3C Working Draft 11 February 2005, http ://www.w3c.org/TR/2005/WD-xpath20-20050211

Bosque, I. & Picallo, C. (1996). Postnominal adjectives in Spanish DPs. In *Journal of Linguistics* 32 (pp. 349-385).

Bosque, I. y Demonte, V. (Eds) (1999). Gramática descriptiva de la lengua española. Esapsa Calpe

Bouillon, P. et al. (1997). Traitement automatique des langues naturelles. Duculot-AUPELF-UREF

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. (Eds) (2004[3]). Extensible Markup Language (XML) 1.0 (third edition) W3C Recommendation 04 February 2004, http://www.w3.org/TR/2004/REC-xml-20040204/

Carlsson, L. (1966). Le degré de cohésion des groupes subst. + de + subst. en français contemporain étudié d'après la place accordée à l'adjectif épithète. Avec examen comparatif des groupes correspondants de l'italien et de l'espagnol. Almqvist & Wikselles

de Schutter, G. en van Hauwermeiren, P. (1983). De Structuur van het Nederlands. Taalbeschouwelijke grammatica. De Sikkel

Derose, S., Maler, E., Orchard, D. (Eds) (2001). XML Linking Language (XLink) 1.0, W3C Recommendation 27/06/2001, http ://www.w3.org/TR/2001/REC-xlink-20010627/

Gazdar, G., Klein, E., Pullum, G. & Sag, I. (1985). Generalized Phrase Structure Grammar. Harvard University Press

Haeseryn, W. (Ed.), Romijn, K., Geerts, G. (1997). Algemene Nederlandse Spraakkunst. M. Nijhoff

Hetzron, R. (1978). On the relative order of adjectives. In H. J. Seiler (Ed.), *Language Universals*, (pp. 165-184) Gunter Narr Verlag,

ISO/IEC (1997). Information technology - Hypermedia/Time based Structuring Language (HyTime) 2d edition, ISO/IEC 10744 : 1997, ISO (International Standardization Organization) and IEC (International Electrotechnical Commission), http ://www.ornl.gov/sgml/wg8/docs/n1920/html/n1920.html

Patota, G. (2003). Grammatica di riferimento della lingua italiana per stranieri. Società Dante Alighieri-Le Monnier

Pollard, C. & Sag, I. A. (1994). Head-Driven Phrase Structure Grammar. CSLI, The University of Chicago Press

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). A comprehensive grammar of the English usage. Longman

Renzi, L. e Salvi, G. (1988). Grande grammatica di consultazione. Il Mulino

Sabarthez, L. (1998). SGML Applications for Linguistic Engineering. In *XML & SGML in action (Proceedings of the SGML BeLux '98 Conference, Antwerpen, 21 October, 1998)*, SGML Belux

Sperberg-McQueen, C. M. & Burnard, L. (Eds) (1994-2002). Guidelines for Electronic Text Encoding and Interchange (TEI P3). http://etext.lib.virginia.edu/teip4/. 15.04.2005

Tennison, J. (2002). Beginning XSLT. Wrox

Wehrli, E. (1997). L'analyse syntaxique des langues naturelles. Problèmes et méthodes. Masson

Wilmet, M. (2003[3]) [1997]. Grammaire critique du français. Duculot