Adapting SimpleNLG for bilingual English-French realisation

Pierre-Luc Vaudry and Guy Lapalme
RALI-DIRO – Université de Montréal
C.P. 6128, Succ. Centre-Ville
Montréal, Québec, Canada, H3C 3J8
{vaudrypl,lapalme}@iro.umontreal.ca

Abstract

This paper describes SimpleNLG-EnFr, an adaption of the English realisation engine SimpleNLG (Gatt and Reiter, 2009) for bilingual English-French realisation. Grammatical similarities between English and French that could be exploited and specifics of French that needed adaptation are discussed.

1 Introduction

Surface realisation is the last step in natural language generation. It takes as input an abstract representation where lexical units and syntactic structures have been determined. Its output is formatted natural language text. SimpleNLG, as described in Gatt and Reiter (2009), is a realisation engine for English in the form of a Java library. It handles inflection, derivation, word order, auxiliaries, agreement, pronominalisation, punctuation, spacing, etc. This paper describes SimpleNLG-EnFr 1.1[1], a bilingual realisation engine for English and French derived from SimpleNLG 4.2, and explains the design choices and the challenges encountered. Grammatical similarities and differences between English and French that influenced the design are discussed. The current version of SimpleNLG is 4.4, but all mentions of SimpleNLG in this paper refer to version 4.2.

2 Subset of French covered

The English grammatical coverage of SimpleNLG-EnFr is the same as that of SimpleNLG 4.2. Its French grammatical coverage is equivalent to its English one.

Le français fondamental (1er Degré) (Ministère de l'Éducation nationale, 1959) was used as a reference for French grammatical coverage. That document results from empirical studies and aims at describing the essential notions for teaching French as a foreign language. Almost all of the grammar points enumerated in this document are covered by SimpleNLG-EnFr. The detailed French grammar rules used in the implementation come mainly from Grevisse (1993) and Mansouri (1996).

SimpleNLG-EnFr has a 3871 entry default French lexicon covering *L'échelle Dubois-Buyse d'orthographe usuelle française* (Ters et al., 1964). It contains the most important and commonly used French vocabulary (including function words), so as not to interfere with a particular application domain vocabulary. A domain specific lexicon can easily be added as SimpleNLG supports using multiple lexicons. Most of the inflected forms in the default French lexicon were taken from Morphalou 2.0 (CNRTL).

3 SimpleNLG parts pooled for English and French

Most of the basic framework, which defined the class hierarchy covering lexical units, phrases and document elements such as paragraphs, could be kept in common for both English and French. Some shared grammar rules and principles were put in abstract classes from which language-specific modules could be derived. The other grammar rules were rewritten for French, with the corresponding English ones serving as references. Many static methods in the English modules in SimpleNLG were changed to regular instance methods in order to be able to override them in the new subclasses.

[1] Available online, along with the source code, at http://www-etud.iro.umontreal.ca/~vaudrypl/snlgbil/snlgEnFr_english.html

Proceedings of the 14th European Workshop on Natural Language Generation, pages 183–187,
Sofia, Bulgaria, August 8-9 2013.

3.1 General characteristics

Features: SimpleNLG uses a system of features for various functions: encoding morphological and syntactic properties of lexical units; letting the user set the parameters of a particular phrase (plural, verb tenses, etc.); and internally keeping track of the content of a phrase and various information needed during realisation. This system is generic enough to be used for other languages. Most features are reusable and others can be added as needed. In SimpleNLG-EnFr, most of the already present features were reused for French.

Lexicon: In SimpleNLG, the lexicon is already relatively well separated from the grammar. The basic lexicon class provides an interface to a simple XML file containing the necessary information about the lexical units. The list of available fields in this file can easily be extended by adding lexical features to the ones used for English. In SimpleNLG-EnFr, many lexical features were added mainly to account for the higher complexity of French morphology.

3.2 Syntax

Verb phrase and clause: First, English and French have the same basic clause constituent order: Subject-Verb-Object (SVO). Even more importantly for SimpleNLG-EnFr, this constituent order is relatively stable (compared with other languages like German or Russian), at least for the purpose of practical NLG applications. This frees us in most cases from having to choose between different syntactically correct word orders. We thus did not have to make such big changes to the syntactic representation as were needed in adapting SimpleNLG to German (Bollmann, 2011). Indeed, in German the subject has the same syntactic status in the clause than the object(s) and they can all occupy the same varying positions relative to the verb. However, Bollmann (2011) had more leeway because he had decided not to keep the English grammar alongside the German one in his implementation. In contrast, in SimpleNLG-EnFr we wanted to be able to change freely between English and French grammars during the generation of a single text.

English and French also have a very similar passive construction. In French, it is used less frequently because other options exist to avoid mentioning the subject of a sentence (for example, using the indefinite personal pronoun *on*), but choosing between those constructions is not the role of the realisation engine.

Noun phrase: English and French can both have a determiner at the beginning of a noun phrase.

Prepositional phrase: Both languages use prepositions (not postpositions) for introducing various complements.

Coordinated phrase: Both have a coordination conjunction in penultimate position and both use commas as separators between coordinates.

3.3 Morphology

In both languages, nouns and verbs are marked morphologically for singular/plural. In addition, personal pronoun forms differ based not only on number and person, but also on grammatical function and gender. This last similarity facilitated adapting pronominalisation.

4 Adaptations for French

The rules for each processing level are encoded in separate modules for each language. The following adaptations were made for French by adding syntactic and lexical features and encoding the corresponding rules in the French versions of the grammar rules modules.

4.1 Syntax

Verb phrase and clause: French negation has some similarities but also big differences with its English counterpart. It is usually expressed with not one but two adverbs (*ne* and *pas*), which come respectively before and after the first word of the verb group, as in example (1). Moreover, *pas* can be replaced by other negation auxiliaries to specify a different kind of negation, as in (2). Finally, no negation auxiliary is used (only *ne*) when the sentence already carries another negative element, for example a negative indefinite pronoun as in (3).

(1) *il ne parle pas*
"he does not speak"

(2) *il ne parle plus*
he not speaks more
"he does not speak anymore"

(3) *personne ne parle*
nobody not speaks
"nobody speaks"

In French, some complement pronouns, instead of being placed after the verb as in the regular SVO word order, are placed just before it. Furthermore, some of them sometimes take in that case a different form. The rules governing

the acceptable combinations and sequencings of those complements that can be cliticised in this way are very precise. Examples (4) and (5) illustrate this phenomenon.

(4) *il la leur réfère*
he her them refers
"he refers her to them"

(5) *il nous réfère à eux*
he us refers to them
"he refers us to them"

The complexity of French past participle agreement is well known, particularly because it manifests itself mostly in written French. French verbs can have *être* (to be) or *avoir* (to have) as auxiliaries in compound tenses. This influences whether the past participle agrees with the subject (*être*) or the direct object if it is placed before the past participle (*avoir*). Combined with clitic complement pronouns and relative clauses, among others, it can get very complex. In addition, French past participles are inflected in gender and number, like adjectives.

Noun phrase: In SimpleNLG, a noun phrase can have pre-modifiers and post-modifiers. Adjectives are by default considered pre-modifiers and everything else post-modifiers. In contrast, in French, most adjectives are placed after the noun, but some (the most common) are most frequently placed before the noun. In SimpleNLG-EnFr this is achieved by referring to an extra lexical feature.

In addition, in French the determiner and adjectives agree with the noun in number and gender. Instead of adding a new mechanism to propagate relevant features of the noun phrase to where they are needed, as with subject-verb agreement in SimpleNLG, the solution implemented was to let the determiner and adjectives get themselves the information they needed from their parent constituent. This more flexible way of managing agreement is more amenable to multilingual realisation.

Interrogative clause: A simple way of building an interrogative sentence in French is to prepend the expression *est-ce que* (is it that), like in (6). This is what we chose.

(6) *est-ce que tu as mangé?*
is it that you have eaten?
"did you eat?"

This kind of interrogative clause can be built in part by using the relative clause rules (see below).

Relative clause: A mechanism for building relative clauses has been added to the French part of SimpleNLG-EnFr that has no direct equivalent in the English implementation. The phrase that must be replaced by a relative pronoun is specified by setting a feature on the clause. This phrase will not appear in the realised clause. Even if this phrase was not present in the clause, it will still be used to choose a relative pronoun, which can be useful. The grammatical function of that phrase can in that case be set manually.

The resulting relative pronoun takes the place that is normally reserved for the complementiser. Its form is chosen according to two sources: the grammatical function and preposition, if any, of the phrase it replaces; and the person and gender of its antecedent (the noun or pronoun that the relative clause modifies). Examples (7), (8) and (9) illustrate this.

(7) *la tarte que tu as mangé**e***
the pie that.obj you have eaten.fem
"the pie that you ate"

(8) *la tarte qui a été mangé**e***
the pie that.subj has been eaten.fem
"the pie that was eaten"

(9) *l'homme dont j'ai mangé la tarte*
the man whose I have eaten the pie
"the man whose pie I ate"

4.2 Morphology

Number and gender: French determiners and adjectives must be inflected in number and gender. Additionally, number and gender interact with each other in the inflection process.

Verb tenses: Verb inflected forms are more varied in French than in English. In addition, French verbs are classified in three conjugation groups. The first group is comprised of the regular verbs. The third group is a catchall category for miscellaneous irregular verbs. Several morphological rules govern the combination of the verb inflection morphemes.

Detached form of personal pronouns: In French, personal pronouns are often cliticised (see subsection 5.1), but where they are not, they take a different form, which is called *forme disjointe* (detached form). See *leur* versus *eux* in examples (4) and (5).

4.3 Morphophonology

The morphophonological level is a new processing level introduced in SimpleNLG-EnFr to account for a range of phenomena very common in French and other languages. They are best described using rules that use both morphological and phonological conditions. The only obvious example of this kind of rule in written English, which was included in the morphology module

in SimpleNLG, is illustrated by examples (10) and (11).

(10) a + book → a book

(11) a + apple → a**n** apple

Here the morphological condition is the presence of the indefinite singular determiner *a* and the phonological one is the presence of a vowel at the beginning of the next word.

The morphology rules operate on one word at a time. The morphophonology rules may need to have access to adjacent words and to be applied after all inflection and derivation rules have been applied. This justifies a separate processing level. In SimpleNLG-EnFr, the morphophonological level is used mainly for external sandhi, i.e. phenomena occurring at word boundaries.

Elision: In French some words have their last vowel elided when in front of a word beginning by a vowel or a so-called *h aspiré* (aspired h). Indeed, an extra lexical feature is needed for French words beginning with the letter *h* to know if that kind of rule applies. Note that the letter *h* itself is never pronounced in French. Examples (12) and (13) illustrate elision, while it does not occur in (14).

(12) *la + amitié → l'amitié*
the friendship

(13) *le + homme → l'homme*
the man

(14) *la + honte → la honte*
the shame

Liaison: *Liaison* is a phenomenon akin to elision, except that it involves adding and/or replacing phonemes. Its "goal" is to avoid contact between the vowel at the end of some words and the beginning vowel of the next word. It is mostly apparent in speech, although it sometimes has an effect in written French, as in (15).

(15) *le + beau + homme → le bel homme*
the handsome man

Prepositions: Some prepositions interact with definite determiners in French, as in (16).

(16) *à + le → au*
at the

5 Bilingual generation

Building a bilingual realisation engine rather than just adapting SimpleNLG for unilingual French realisation was a design choice dictated mainly by practical considerations. Being able to use the same realisation engine (and thus the same API) for several or all target languages when developing a multilingual NLG application is convenient. In the case of English and French, this could be most useful when targeting Canadian or European populations, for example.

In SimpleNLG-EnFr, bilingual generation is implemented by being able to determine dynamically the language of each processing unit: phrases for the syntax module, lexical units for the morphology module, etc. The factories used by the library's user to create syntactic structure specifications and access or create lexical units each use a language-specific lexicon. Each processing module then chooses at realisation time which set of rules to apply to a given processing unit based on the language of its lexicon. Thus, sentences, phrases and words of different languages can be mixed freely.

6 Conclusion

A bilingual realisation engine for English and French was built. It took five months to complete, including the writing of a detailed French manual. Despite many internal changes, it retains almost the same API as the original.

Future improvements could include enlarging the default lexicon and adding specialised lexicons for French, implementing a complete textual representation for numbers, and adapting the changes in SimpleNLG since version 4.2, like the XML realiser.

More languages could be added to SimpleNLG-EnFr. However, it would perhaps be easier to include many languages if the grammar of each language could be specified in a common grammar formalism, instead of programmatically in the processing modules themselves. This would necessitate changing the architecture.

In the process of developing SimpleNLG-EnFr, a great deal was learned about what kind of challenges multilingual realisation poses. A common grammatical ground must be found and exploited for the group of languages considered, which should not be too far apart in that respect. For the rest, care must be taken not to make too many assumptions about the inner workings of the grammar of each language. Indeed, every language has its own grammatical peculiarities.

Acknowledgments

We would like to thank the SimpleNLG team for having made available its source code. This work was supported by two Undergraduate Student Research Awards (USRA) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Marcel Bollmann. 2011. Adapting SimpleNLG to German. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011)*, pages 133-138.

CNRTL. *CNRTL : Centre National de Ressources Textuelles et Lexicales – Morphalou,* button *Télécharger Morphalou 2.0*, [http://www.cnrtl.fr/lexiques/morphalou/] (consulted on 14 July 2011).

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90-93.

Maurice Grevisse, (1993). *Le bon usage, grammaire française*, 12e édition refondue par André Goosse, 8e tirage, Éditions Duculot, Louvain-la-Neuve, Belgique.

Mohammed Issaoui Mansouri, (1996). *Le Mansouris, tous les verbes usuels entièrement conjugués et orthographiés*. CAPT, Éditeurs, Montréal, Canada.

Ministère de l'Éducation nationale, Direction de la Coopération avec la Communauté et l'Étranger (France) (1959). *Le français fondamental (1er Degré)*, Publication de l'Institut Pédagogique National, Paris, France.

François Ters, Daniel Reichenbach and Georges Mayer. (1964). *L'échelle Dubois-Buyse d'orthographe usuelle française*. Messeiller.