Lexical Resources for Automatic Translation of Constructed Neologisms: the Case Study of Relational Adjectives

Bruno Cartoni
ISCCO/TIM/ETI- University of Geneva
40 bd du Pont-d'Arve, CH-1205 Geneva
E-mail: cartoni5@etu.unige.ch

Abstract

This paper deals with the treatment of constructed neologisms in a machine translation system. It focuses on a particular issue in Romance languages: relational adjectives and the role they play in prefixation. Relational adjectives are formally adjectives but are semantically linked to their base-noun. In prefixation processes, the prefix is formally attached to the adjective, but its semantic value(s) is applied to the semantic features of the base-noun. This phenomenon has to be taken into account by any morphological analyser or generator. Moreover, in a contrastive perspective, the possibilities of creating adjectives out of nouns are not the same in every language. We present the special mechanism we put in place to deal with this type of prefixation, and the automatic method we used to extend lexicons, so that they can retrieve the base-nouns of prefixed relational adjectives, and improve the translation quality.

1. Introduction

Within machine translation systems that deal with constructed words, simple decomposition in one language and mechanical reconstruction in another one are rarely efficient enough to provide a correct translation. Once the morphological analysis of the constructed neologism has succeeded, (i.e. the neologism has been identified as such and not confused with a homographic form – proper noun, misspelling, …), there remain some morphological phenomena to deal with that require particular lexical and translation resources. In this study, we show the benefit brought by the extension of a lexicon with relational adjectives, especially in the translation of prefixed Italian neologisms into French. We first explain the general principles of our translation system, focusing on the treatment we propose for the prefixation processes on relational adjectives, and then we explain how we created special resources to deal with relational adjectives and evaluate the benefit of including them into a morphology-based automatic translation system.

2. Description of the system

Neologisms are problematic for NLP systems, and especially for machine translation systems, because neologisms are not analysed, and not translated (Gdaniec, Manandise *et al.* 2001). The study presented here is performed in the framework of a experimental system that translates constructed neologisms from Italian into French. This system is composed of two modules. The first one checks every unknown word to see if it is potentially constructed. The second module is the actual translation module, which analyses the constructed neologism and generates a possible translation. The first module has already been evaluated and produced satisfying results (Cartoni 2006; Cartoni 2007). We focus here on the second module, and especially on the use and the implementation of special lexical resources.

The translation of neologisms relies on the presupposition that morphological processes can be transfered from one language to another. So, for a constructed neologism in one language (e.g. *ricostruire* in Italian), the system makes a morphological analysis to find the rule that produced the neologism (in this case *ri+costruire* <reiteration rule>), and then, through a transfer mechanism, generates a translation, either by rebuilding a constructed word, (*reconstruire*, to rebuild) or by proposing a paraphrase (*construire à nouveau*, to build again). The whole process is formalised into bilingual Word Formation Rules (WFR), such as the one shown in Figure 1 for reiterativity prefixation. The first line is the centre of the rule, describing the production of a verb (y_V) using a base verb (x_V) and a prefix (*ri* or *re*). The next line states a constraint put on the base (here, being in the reference monolingual lexicon). This constraint might seem very strict, but avoids a lot of noise in the analysis of unknown words that begin with *ri* and that are not constructed neologisms. Finally, the last line contains semantic information and/or a « paraphrase » that can be used as an alternative translation.

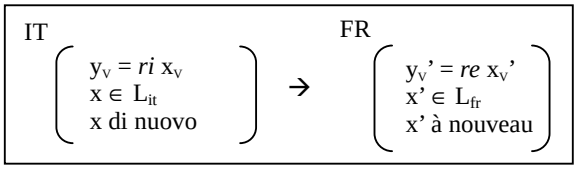

Figure 1: Bilingual WFR for reiterativity

From the lexical point of view, our prototype is based on two very large monolingual databases (Mmorph (Bouillon, Lehmann *et al.* 1998)) and a semi-automatically constructed bilingual lexicon, which matches together the two monolingual database. This bilingual lexicon is very small, and built from scratch to meet the needs of the experiment described here.

3. Problems in translating the base: the relational adjective

Translating a prefixed word does not mean concatenating the translation of the prefix with the translation of the base, especially because the semantic base of prefixed adjective sometime does not correspond to the formal base. This happens for a very common phenomenon in Romance languages: the prefixation of *relational adjectives.* Relational adjectives are derived from nouns and designate a relation between the entity denoted by the noun they are derived from and the entity denoted by the noun they modify.

Consequently, in a prefixation such as *anticostituzionale*, the formal base is a relational adjective (*costituzionale*), but the semantic base is the noun the adjective is derived from (*costituzione*). The constructed word *anticostituzionale* can be paraphrased as "*against the constitution*". Moreover, when the relational adjective does not exist, prefixation is possible on a nominal base to create an adjective (*squadra antidroga*). In cases where the adjective does exist, both forms are possible and seem to be equally used, like in the Italian *collaborazione interuniversità / collaborazione interuniversitaria*.

From a contrastive point of view, the prefixation of relational adjective exists in both languages (Italian and French) and in both these languages prefixing a noun to create an adjective is also possible (*anticostituzione* (Adj)). But we observe an important discrepancy in the possibility of constructing relational adjectives, as shown in the evaluation summarised below.

3.1 Divergence between languages in constructing relational adjectives

A small experiment based on the Italian-French Garzanti dictionary (2006) shows that adjectival denominalisation (i.e the process that makes an adjective out of a noun) is very different in the French and Italian languages.

Of a total of more than 10'000 Italian adjectives, a rough estimation shows that about 1'000 adjectives have no adjectival French equivalents. In the dictionary, they are generally translated by a prepositional phrase containing the base noun, like in the examples shown below:

adolescenziale → de l'adolescence
aziendale → de l'entreprise
creditizio → de crédit
gattesco → de chat
partitico → de parti
congressuale → du congrès

If one of these relational adjectives is used in a prefixation process (like in *precongressuale*), the translation mechanism has to find the base noun of the adjective (*congresso →congressuale*) in order to be able to generate in French a constructed neologism (*précongrès*) or a phrase (*avant le congrès*).

3.2 Proposed solution

To deal with the prefixation on relational adjectives and the discrepancy between the two languages, we propose to implement bilingual WFR in order to take into account this phenomenon, as shown in figure 2 for the WFR for the opposition in *anti.*

In this rule, the base is analysed to find the base noun of the relational adjective ($[\ (z')_N\ ]_{REL\_ADJ}$), and semantic instructions are applied on the base noun (contro z).

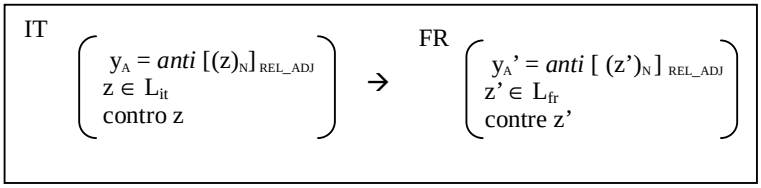

Figure 2 : Bilingual WFR for opposition in *anti*

Taking this phenomenon into account is very useful for many aspects: (1) the analysis quality is much more detailed, (2) the information can be used to generate a paraphrase, in Italian or as a translation in French, and (3) it gives the possibility of translating/generating a noun-based prefixed adjective (like *antidroga*), which is especially useful if the relational adjective is not available in the target language, or if it is simply missing in the system lexicon.

But, theses rules require appropriate lexical resources. In the following sections, we sketch out the resources, present a way to acquire them, and evaluate their benefit.

4. Extending lexical resources to deal with relational adjectives

Our system is based on a reference lexicon for Italian ("L_{it}" in the rules shown above) that provides morphosyntactic information for the base word, but not information on relational adjectives, as explained above. Consequently, we looked for a simple way to automatically extend the Italian lexicon so that it could make the link between a relational adjective and its noun base, and provide this information during the analysis process.

Some projects have already dealt with this issue, but mainly by acquiring relational adjective from corpora (e.g. (Daille 1999)). Our approach, on the other hand, tries to take advantage of only the lexicon, without the use of any larger resources. To extend the Italian lexicon, we simply built a routine based on the typical suffixes of relational adjectives (in

Italian: *-ale, -are, -ario, -ano, -ico, -ile, -ino, -ivo, -or io, -esco, -asco, -iero, -izio, -aceo* (Wandruszka 2004))
For every adjective ending with one of these suffixes, the routine looks up if the potential base corresponds to a noun in the rest of the lexicon (modulo some morphographemic variations). For example, the routine is able to find links between adjectives and base nouns such as *ambientale* and *ambiente*, *aziendale* and *azienda*, *cortisonica* and *cortisone* or *contestuale* and *contesto.*
Unfortunately, this kind of automatic implementation does not find links between adjectives made from the learned root of the noun, (*prandiale* → *pranzo, bellico* → *guerra*). This lack is probably the cause for the low recall of this automatic extension. But, results are much better than expected regarding the precision, as we show below, in the qualitative evaluation of the extension.

4.1 Evaluation of the extended lexical resources

We evaluated for every suffix the number of wrong links between one adjective and one noun, and kept only the suffixes that guaranteed a precision above 90%, in order to get a relational adjective lexicon as precise as possible. Consequently, we excluded the suffixes: *-ile* (precision: 53%), *-ano* (54%), *-iano* (46%), and *–iario* (48%).
With the remaining rules, and from a total of more than 68'000 adjective forms in the lexicon, we identified 8'466 relational adjectives. From a "recall" perspective, it is not easy to evaluate the coverage of this extension because of the small number of resources containing relational adjectives that could be used as a gold standard. But we can estimate that a majority are qualification adjectives.
Another way to evaluate the quality of this extension is to measure the improvement brought by it to the translation process. This is what we propose in the following section.

5. Integrating the rules into the system

We include this extended lexicon in the translation module of the proposed system and adapt prefixation rules consequently. This phenomenon is actually applicable to different classes of prefixes: the quantitative prefixes (*pluri, poli, tri, uni. mono, multi bi, di*), the locating prefixes (*neo, oltre, para, ex, extra, inter, intra, meta, post, pre, pro, sopra, sovra, sotto, sub, super, trans*), and some negative prefixes (*a, anti*).
Figure 3 below shows the mechanism and the many possible translations that these implemented rules make possible. When an Italian constructed neologism arrives into the system (here: *anticostituzionale*), it is analysed by the rule shown in Figure 2, and the formal base (i.e the adjective) is looked up in the bilingual lexicon (step 1). If this base is recorded in the lexicon, the neologism can be easily generated in French. If not, the adjective-base is looked up in the monolingual Italian lexicon to find the nominal base (*costituzione*) (step 2). This nominal base is then found in the bilingual dictionary (step 3). Then, two options are possible. Either the translation is generated on a nominal base (step 4, *anticonstitution*) or the French relational adjective is found in the French monolingual lexicon (step 5 *constitution* → *constitutionel*) and the neologism is generated in French (step 6 : *anticonstitutionnel*).
In some cases, the extended system and lexicon has allowed for the proposal of a translation with a nominal base when the relational adjective was not in the bilingual dictionary. For example, Italian *antileucemico* is constructed from the relational adjective *leucemico* which derives from the noun *leucemia.* The bilingual lexicon does not contain an entry for *leucemico*, only an entry for the noun (*leucemia=leucémie*). Thanks to the

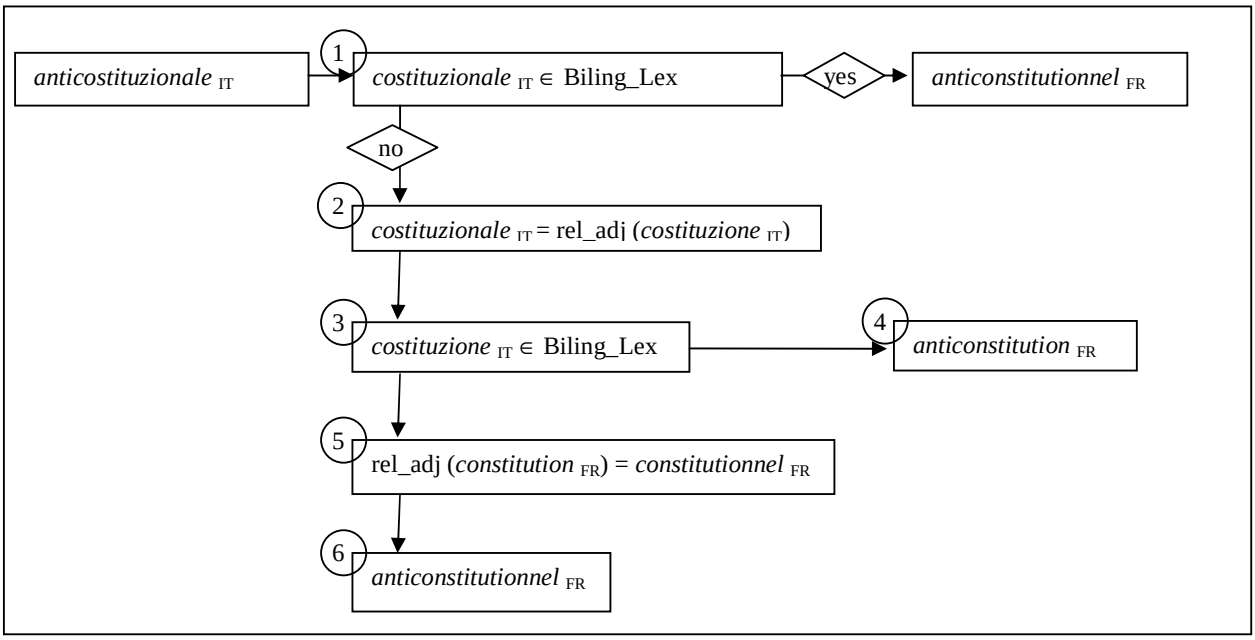

Figure 3 : Mechanism for translating with different bases

extended lexicon and the fine-grained information that links the adjective *leucemico* with the noun *leucemia*, the system can generate a French translation using the French noun base (*antileucémie*).

6. Evaluation of translation

To evaluate this system globally, we extracted a set of 24'247 unknown words from the corpus *La Reppublica* (Baroni, Bernardini *et al.* 2004), that were potential prefixed neologisms. The translation system with no extension of the lexicon with relational adjectives translated 17034 neologisms (68,76 %). Amongst these 17'034 neologisms, 5'025 are constructed with the 28 prefixes which might have a relational adjective as a base. And amongst them, the extended lexicon is able to identify 1'783 relational adjectives, which is an important improvement in terms of the quality of the analysis. For example, thanks to the extended resources, the analysis now provides a mechanical decomposition of the constructed neologism together with the base noun of the relational adjective, like (e.g. *multidisciplinare* → *multi*disciplinare*$_{A}$ /*disciplina*$_{N}$, *sottoministeriali* → *sotto*ministeriali*$_{A}$ /*ministero*$_{N}$, *antidemocratico* → *anti*democratico*$_{A}$ /*democrazia*/$_{N.}$). On the generation/translation side, all neologisms have been translated, the majority (1'570) by a prefixed relational adjective and the rest (213) by a French noun, because the relational adjective was not in the bilingual lexicon. And, amongst this last group, we found interesting cases where the lack of the French relationnal adjective is not only a lack in the bilingual lexicon, but a non-existant word in the French language, such as *precongressuale* → *précongrès*, *post-transfuzionale* → *post-transfusion*, *predibatimentale* → *prédébat*). Particularly for these last cases, a translation using simple decomposition and reconstruction would give no results.

So, the extension of the lexicon has two advantages. First, the relational adjectives are better analyzed, and second, when the adjectival base is not in the bilingual lexicon, the translation can never the less be done.

7. Conclusion and ongoing work

This preliminary study shows the possible improvement gained through the use of relational adjectives for translating constructed words. Thanks to the extended resources, we increase the number of words translated correctly. Indeed, the "non-translation" of constructed words is typically due to the lack of the base word in the lexicon. Finding the nominal base of a relational adjective is consequently a good solution for solving this problem.

Further work is currently being done to (1) extend the French lexicon with the same kind of links, in order to generate the relational adjective from the noun in the target language, (2) add links between geographical nouns and their relational adjectives and (3) evaluate from a qualitative perspective the output of the translation. Finally (4), we are currently assessing the possibility of exploiting other links within the lexicon, such as for deverbal nouns or adjectives, for which the prefixation is applied on the verbal base of the formal base (like in *anticoagulation* → 'that prevents to coagulate').

The experiment presented here also allows us to imagine that bilingual resources might not need to be extended as much if monolingual relational links are provided. But, we also believe that extending a lexicon with this kind of information could be exploited for other purposes, beyond its application to constructed neologisms. For example, it is well known that Germanic languages tend to prefer compounding N+N (e.g. English: *muscle fiber*) where Romance languages prefer the structure N+Adj_rel (e.g Italian: *fibra muscolare*). Linking a noun and a relational adjective (*muscolare* → *muscolo* → *muscel*) in a multilingual perspective would probably benefit the quality of machine translation.

8. References

(2006) Garzanti francese: francese-italiano, italiano-francese. I grandi dizionari Garzanti. Milano, Garzanti Linguistica.

Baroni, M., S. Bernardini, *et al.* (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. Proceedings of LREC 2004., Lisbon.

Bouillon, P., S. Lehmann, *et al.* (1998). Développement de lexiques à grande échelle. Colloque des journées LTT de TUNIS, Tunis.

Cartoni, B. (2006). Dealing with unknown words by simple decomposition: feasibility studies with Italian prefixes. LREC 2006, Gênes.

Cartoni, B. (2007). Régler les règles d'analyse morphologique. TALN 2007, Toulouse, IRIT.

Daille, B. (1999). Identification des adjectifs relationnels en corpus. Conference TALN 1999, Cargèse.

Gdaniec, C., E. Manandise, *et al.* (2001). Derivational Morphology to the Rescue: How It Can Help Resolve Unfound Words in MT. MT Summit VIII, Santiago Di Compostella.

Iacobini, C. (2004). I prefissi. La formazione delle parole in italiano. M. Grossmann et F. Rainer. Tübingen, Niemeyer**:** 99-163.

Wandruszka, U. (2004). Derivazione aggettivale. La Formazione delle Parole in Italiano. M. Grossman et F. Rainer. Tübingen, Niemeyer.