Lexical Morphology in Machine Translation: a Feasibility Study

Bruno Cartoni
University of Geneva
cartonib@gmail.com

Abstract

This paper presents a feasibility study for implementing lexical morphology principles in a machine translation system in order to solve unknown words. Multilingual symbolic treatment of word-formation is seducing but requires an in-depth analysis of every step that has to be performed. The construction of a prototype is firstly presented, highlighting the methodological issues of such approach. Secondly, an evaluation is performed on a large set of data, showing the benefits and the limits of such approach.

1 Introduction

Formalising morphological information to deal with morphologically constructed unknown words in machine translation seems attractive, but raises many questions about the resources and the prerequisites (both theoretical and practical) that would make such symbolic treatment efficient and feasible. In this paper, we describe the prototype we built to evaluate the feasibility of such approach. We focus on the knowledge required to build such system and on its evaluation. First, we delimit the issue of neologisms amongst the other unknown words (section 2), and we present the few related work done in NLP research (section 3). We then explain why implementing morphology in the context of machine translation (MT) is a real challenge and what kind of aspects need to be taken into account (section 4), and we show that translating constructed neologisms is not only a mechanical decomposition but requires more fine-grained analysis. We then describe the methodology developed to build up a prototyped *translator* of constructed neologisms (section 5) with all the extensions that have to be made, especially in terms of resources. Finally, we concentrate on the evaluation of each step of the process and on the global evaluation of the entire approach (section 6). This last evaluation highlights a set of methodological criteria that are needed to exploit lexical morphology in machine translation.

2 Issues

Unknown words are a problematic issue in any NLP tool. Depending on the studies (Ren and Perrault 1992; Maurel 2004), it is estimated that between 5 and 10 % of the words of a text written in “standard” language are unknown to lexical resources. In a MT context (analysis-transfer-generation), unknown words remain not only unanalysed but they cannot be translated, and sometimes they also stop the translation of the whole sentence.

Usually, three main groups of unknown words are distinguished: proper names, errors, and neologisms, and the possible solution highly depends on the type of unknown word to be solved. In this paper, we concentrate on neologisms which are constructed following a morphological process.

The processing of unknown “constructed neologisms” in NLP can be done by simple guessing (based on the sequence of final letters). This option can be efficient enough when the task is only tagging, but in a multilingual context (like in MT), dealing with constructed neologisms implies a transfer and a generation process that require a more complex formalisation and implementation. In the project presented in this paper, we propose to implement lexical morphology phenomena in MT.

3 Related work

Implementing lexical morphology in a MT context has seldom been investigated in the past, probably because many researchers share the following view: “Though the idea of providing rules for translating derived words may seem attractive, it raises many problems and so it is currently more of a research goal for MT than a practical possibility” (Arnold, Balkan et al. 1994). As far as we know, the only related project is described in (Gdaniec, Manandise et al. 2001), where they describe a project of implementation of rules for dealing with constructed words in the IBM MT system.

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 130–138,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

Even in monolingual contexts, lexical morphology is not very often implemented in NLP. Morphological analyzers like the ones described in (Porter 1980; Byrd 1983; Byrd, Klavans et al. 1989; Namer 2005) propose more or less deeper lexical analyses, to exploit that dimension of the lexicon.

4 Proposed solution

Since morphological processes are regular and exist in many languages, we propose an approach where constructed neologisms in source language (SL) can be analysed and their translation generated in a target language (TL) through the transfer of the constructional information.

For example, a constructed neologism in one language (e.g. *ricostruire* in Italian) should firstly be analysed, i.e. find (i) the rule that produced it (in this case <reiteration rule>) and (ii) the lexeme-base which it is constructed on (*costruire,* with all morphosyntactic and translational information). Secondly, through a transfer mechanism (of both the rule and the base), a translation can be generated by rebuilding a constructed word, (in French *reconstruire,* Eng: to rebuild). On a theoretical side, the whole process is formalised into bilingual Lexeme Formation Rules (LFR), as explained below in section 4.3.

Although this approach seems to be simple and attractive, feasibility studies and evaluation should be carefully performed. To do so, we built a system to translate neologisms from one language into another. In order to delimit the project and to concentrate on methodological issues, we focused on the prefixation process and on two related languages (Italian and French). Prefixation is, after suffixation, the most productive process of neologism, and prefixes can be more easily processed in terms of character strings. Regarding the language, we choose to deal with the translation of Italian constructed neologisms into French. These two languages are historically and morphologically related and are consequently more "neighbours" in terms of neologism coinage.

In the following, we firstly describe precisely the phenomena that have to be formalized and then the prototype built up for the experiment.

4.1 Phenomena to be formalized

Like in any MT project, the formalisation work has to face different issues of contrastivity, i.e. highlighting the divergences and the similarities between the two languages.

In the two languages chosen for the experiment, few divergences were found in the way they construct prefixed neologisms. However, in some cases, although the morphosemantic process is similar, the item used to build it up (i.e. the affixes) is not always the same. For example, to coin nouns of the spatial location "before", where Italian uses the prefix *retro,* French uses *rétro* and *arrière*. A deeper analysis shows that Italian *retro* is used with all types of nouns, whereas in French, *rétro* only forms processual nouns (derived from verbs, like *rétrovision*, *rétroprojection*). For the other type of nouns (generally locative nouns), *arrière* is used (*arrière-cabine*, *arrière-cour).*

Other problematic issues appear when there is more than one prefix for the same LFR. For example, the rule for "indeterminate plurality" provides in both languages a set of two prefixes (*multi/pluri* in Italian and *multi/pluri* in French) with no known restrictions for selecting one or the other (e.g. both *pluridimensionnel* and *multidimensionnel* are acceptable in French). For these cases, further empirical research have to be performed to identify restrictions on the rule.

Another important divergence is found in 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 adjectives 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 notice an important discrepancy in the possibility of constructing relational adjectives (a rough estimation performed on a large bilingual dictionary (Garzanti IT-FR (2006)) shows that more than 1 000 Italian relational adjectives have no equivalent in French (and are generally translated with a prepositional phrase).

All these divergences require an in-dept analysis but can be overcome only if the formalism and the implementation process are done following a rigorous methodology.

4.2 The prototype

In order to evaluate the approach described above and to concretely investigate the ins and outs of such implementation, we built up a prototype of a machine translation system specialized for constructed neologisms. This prototype is composed of two modules. The first one checks every unknown word to see if it is potentially constructed, and if so, performs a morphological analysis to individualise the lexeme-base and the rule that coined it. The second module is the actual translation module, which analyses the constructed neologism and generates a possible translation.

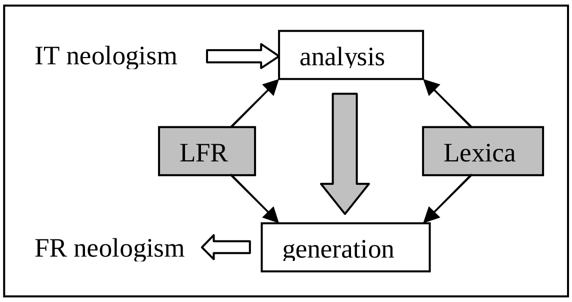

Figure 1: Prototype

The whole prototype relies on one hand on lexical resources (two monolingual and one bilingual) and on a set of bilingual Lexeme Formation Rules (LFR). These two sets of information helps the analysis and the generation steps. When a neologism is looked-up, the system checks if it is constructed with one of the LFRs and if the lexeme-base is in the lexicon. If it is the case, the transfer brings the relevant morphological and lexical information in the target language. The generation step constructs the translation equivalent, using the information provided by the LFR and the lexical resources. Consequently, the whole system relies on the quality of both the lexical resources and the LFR.

4.3 Bilingual Lexeme Formation Rules

The whole morphological process in the system is formalised through bilingual Lexeme Formation Rules. Their representation is inspired by (Fradin 2003) as shown in figure 2 in the rule of reiterativity.

Such rules match together two monolingual rules (to be read in columns). Each monolingual rule describes a process that applies a series of instructions on the different sections of the lexeme : the surface section (G and F), the syntactic category (SX) and the semantic (S) sections. In this theoretical framework, affixation is only one of the instructions of the rule (the graphemic and phonological modification), and consequently, affixes are called “exponent” of the rule.

	Italian		*French*
	input		input
(G)	V_{it}		V_{fr}
(F)	/V_{it}/		/V_{fr}/
(SX)	cat :v		cat :v
(S)	V_{it}'(...)		V_{fr}'(...)
	↕	↔	↕
	output		output
(G)	riV_{it}		reV_{fr}
(F)	/ri/⊕/V_{it}/		/ʀə/⊕/V_{fr}/
(SX)	cat :v		cat :v
(S)	reiterativity (V_{it}'(...))		reiterativity (V_{fr}'(...))

where V_{it}' = V_{fr}', translation equivalent

Figure 2: Bilingual LFR of reiterativity

This formalisation is particularly useful in a bilingual context for rules that have more than one prefix in both languages: more than one affix can be declared in one single rule, the selection being made according to different constraints or restrictions. For example, the rule for “indeterminate plurality” explained in section 4.1 can be formalised as follows:

	Italian		*French*
	input		input
(G)	X_{it}		X_{fr}
(F)	/X_{it}/		/X_{fr}/
(SX)	cat :n		cat :n
(S)	X_{it}'(...)		X_{fr}'(...)
	↕	↔	↕
	output		output
(G)	multi/pluriX_{it}		multi/pluriX_{fr}
(F)	/multi/pluri/⊕/X_{it}/		/mʏlti/plyri/⊕/X_{fr}/
(SX)	cat :n		cat :n
(S)	indet. plur. (X_{it}'(...))		indet. plur. (X_{fr}'(...))

where X_{it}' = X_{fr}', translation equivalent

Figure 3: Bilingual LFR of indeterminate plurality

In this kind of rules with “multiple exponents”, the two possible prefixes are declared in the surface section (G and F). The selection is a monolingual issue and cannot be done at the theoretical level.

Such rules have been formalised and implemented for the 56 productive prefixes of Italian (Iacobini 2004)[1], with their French translation equivalent. However, finding the translation equivalent for each rule requires specific studies

[1] i.e. *a, ad, anti, arci, auto, co, contro, de, dis, ex, extra, in, inter, intra, iper, ipo, macro, maxi, mega, meta, micro, mini, multi, neo, non, oltre, onni, para, pluri, poli, post, pre, pro, retro, ri, s, semi, sopra, sotto, sovra, stra, sub, super, trans, ultra, vice, mono, uni, bi, di, tri, quasi, pseudo.*

of the morphological system of both languages in a contrastive perspective.

The following section briefly summarises the contrastive analysis that has been performed to acquire this type of contrastive knowledge.

4.4 Knowledge acquisition of bilingual LFR

As in any MT system, the acquisition of bilingual knowledge is an important issue. In morphology, the method should be particularly accurate to prevent any methodological bias. To formalise translation rules for prefixed neologisms, we adopt a meaning-to-form approach, i.e. discovering how a constructed meaning is morphologically realised in two languages.

We build up a *tertium comparationis* (a neutral platform, see (James 1980) for details) that constitute a semantic typology of prefixation processes. This typology aims to be universal and therefore applicable to all the languages concerned. On a practical point of view, the typology has been built up by summing up various descriptions of prefixation in various languages (Montermini 2002; Iacobini 2004; Amiot 2005). We end up with six main classes*: location*, *evaluation*, *quantitative, modality*, *negation and ingressive*. The classes are then subdivided according to sub-meanings: for example, *location* is subdivided in *temporal* and *spatial*, and within *spatial location,* a distinction is made between different positions (*before*, *above*, *below*, *in front*, …).

Prefixes of both languages are then literally "projected" (or classified) onto the *tertium*. For each terminal sub-class, we have a clear picture of the prefixes involved in both languages. For example, the LFR presented in figure 1 is the result of the projection of the Italian prefix (*ri*) and the French one (*re*) on the sub-class *reiteratvity,* which is a sub-class of *modality.*

At the end of the comparison, we end up with more than 100 LFRs (one rule can be reiterated according the different input and output categories). From a computing point of view, constraints have to be specified and the lexicon has to be adapted consequently.

5 Implementation

Implementation of the LFR is set up as a database, from where the program takes the information to perform the analysis, the transfer and the generation of the neologisms. In our approach, LFRs are simply declared in a tab format database, easily accessible and modifiable by the user, as shown below:

```
arci  a       a  2.1.2   archi
arci  n       n  2.1.2   archi
[…]
pro   a_rel   a  1.1.10  pro
pro   n       a  1.1.10  pro
[…]
ri    v       v  6.1     re
ri    n_dev   n  6.1     re
[…]
```

Figure 4: Implemented LFRs

Implemented LFRs describe (i) the surface form of the Italian prefix to be analysed, (ii) the category of the base, (iii) the category of the derived lexeme (*the output*), (iv) a reference to the rule implied and (v) the French prefix(es) for the generation.

The surface form in (i) should sometimes take into account the different allomorphs of one prefix. Consequently, the rule has to be reiterated in order to be able to recognize any forms (e.g. the prefix *in* has different forms according to the initial letter of the base, and four rules have to be implemented for the four allomorphs (*in, il, im, ir*)). In some other cases, the initial consonant is doubled, and the algorithm has to take this phenomenon into account.

In (ii), the information of the category of the base has been "overspecified", to differentiate qualitative and relational adjectives, and deverbal nouns and the other ones (`a_rel/a` or `n_dev/n`). These overspecifications have two objectives: optimizing the analysis performance (reducing the noise of homographic character strings that look like constructed neologisms but that are only misspellings - see below in the evaluation section), and refining the analysis, i.e. selecting the appropriate LFR and, consequently, the appropriate translation.

To identify relational adjectives and deverbal nouns, the monolingual lexicon that supports the analysis step has to be extended. Thereafter, we present the symbolic method we used to perform such extension.

5.1 Extension of the monolingual lexicon

Our MT prototype relies on lexical resources: it aims at dealing with unknown words that are not in a Reference lexicon and these unknown words are analyzed with lexical material that is in this lexicon.

From a practical point of view, our prototype is based on two very large monolingual data-bases (*Mmorph* (Bouillon, Lehmann *et al.* 1998)) for Italian and French, that contain only morphosyntactic information, and on one bilingual lexicon that has been built semi-automatically for the use of the experiment. But the monolingual lexica have to be adapted to provide specific information necessary for dealing with morphological process.

As stated above, identifying the prefix and the base is not enough to provide a proper analysis of constructed neologisms which is detailed enough to be translated. The main information that is essential for the achievement of the process is the category of the base, which has to be sometimes "overspecified". Obviously, the Italian reference lexicon does not contain such information. Consequently, we looked for a simple way to automatically extend the Italian lexicon. For example, we looked for a way to automatically link **relational adjectives** with their noun bases.

Our approach 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, -orio, -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 automatic extension has been evaluated. Out of 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.

A similar extension is performed for the deverbal aspect, for the lexicon should also distinguish **deverbal noun**. From a morphological point of view, deverbalisation can be done trough two main productive processes: conversion (*a command* → *to command*) and suffixation. If the first one is relatively difficult to implement, the second one can be easily captured using the typical suffixes of such processes. Consequently, we considere that any noun ending with suffixes like *ione, aggio,*or *mento* are deverbal.

Thanks to this extended lexicon, overspecified input categories (like `a_rel` for *relational adjective* or `n_dev` for *deverbal noun*) can be stated and exploited in the implemented LFR as shown in figure 4.

5.2 Applying LFRs to translate neologisms

Once the prototyped MT system was built and the lexicon adapted, it was applied to a set of neologisms (see section 6 for details). For example, unknown Italian neologisms such as *arcicontento, ridescrizione, deitalianizzare,* were automatically translated in French: *archi-content, redescription, désitalianiser.*

The divergences existing in the LFR of <locative position before> are correctly dealt with, thanks to the correct analysis of the base. For example, in the neologism *retrobottega,* the lexeme-base is correctly identified as a locative noun, and the French equivalent is constructed with the appropriate prefix (*arrière-boutique*), while in *retrodiffusione,* the base is analysed as *deverbal,* and the French equivalent is correctly generated (*rétrodiffusion*).

For the analysis of relational adjectives, the overspecification of the LFRs and the extension of the lexicon are particularly useful when there is no French equivalent for Italian relational adjectives because the corresponding construction is not possible in the French morphological system. For example, the Italian relational adjective *aziendale* (from the noun *azienda,* Eng: company) has no adjectival equivalent in French. The Italian prefixed adjective *interaziendale* can only be translated in French by using a noun as the base (*interentreprise*). This translation equivalent can be found only if the base noun of the Italian adjective is found (*interaziendale, inter+aziendale* → *azienda, azienda = entreprise,* → *interentreprise*). The same process has been applied for the translation of *precongressuale, post-transfuzionale* by *précongrès, post-transfusion.*

Obviously, all the mechanisms formalised in this prototype should be carefully evaluated.

6 Evaluation

The advantages of this approach should be carefully evaluated from two points of view: the evaluation of the performance of each step and of the feasibility and portability of the system.

6.1 corpus

As previously stated, the system is intended to solve neologisms that are unknown from a lexicon with LFRs that exploit information contained in the lexicon. To evaluate the performance of our system, we built up a corpus of unknown words by confronting a large Italian corpus from journalistic domain (*La Repubblica Online* (Baroni, Bernardini et al. 2004*)*) with our reference lexicon for this language (see section 4.1 above). We obtained a set of unknown words that contains neologisms, but also proper names and erroneous items. This set is submitted to the various steps of the system, where constructed neologisms are recognised, analysed and translated.

6.2 Evaluation of the performance of the analysis

As we previously stated, the analysis step can actually be divided into two tasks. First of all, the program has to identify, among the unknown words, which of them are morphologically constructed (and so analysable by the LFRs); secondly, the program has to analyse the constructed neologisms, i.e matching them with the correct LFRs and isolating the correct base-words.

For the first task, we obtain a list of 42 673 potential constructed neologisms. Amongst those, there are a number of erroneous words that are homographic to a constructed neologism. For example, the item *progesso,* a misspelling of *progresso* (Eng: *progress*), is erroneously analysed as the prefixation of *gesso* (eng: *plaster*) with the LFR in *pro.*

In the second part of the processing, LFRs are concretely applied to the potential neologisms (i.e. constraints on categories and on overspecified category, phonological constraints). This stage retains 30 376 neologisms. A manual evaluation is then performed on these outputs. Globally, 71.18 % of the analysed words are actually neologisms. But the performance is not the same for every rule. Most of them are very efficient: among all the rules for the 56 Italian prefixes, only 7 cause too many erroneous analyses, and should be excluded - mainly rules with very short prefixes (like *a, di, s*), that cause mistakes due to homograph.

As explained above, some of the rules are strongly specified, (i.e. very constrained), so we also evaluate the consequence of some constraints, not only in terms of improved performance but also in terms of loss of information. Indeed, some of the constraints specified in the rule exclude some neologisms (false negatives). For example, the modality LFRs with *co* and *ri* have been overspecified, requiring deverbal base-noun (and not just a noun). Adding this constraint improves the performance of the analysis (i.e. the number of correct lexemes analysed), respectively from 69.48 % to 96 % and from 91.21 % to 99.65 %. Obviously, the number of false negatives (i.e. correct neologisms excluded by the constraint) is very large (between 50 % and 75 % of the excluded items).

In this situation, the question is to decide whether the gain obtained by the constraints (the improved performance) is more important than the un-analysed items. In this context, we prefer to keep the more constrained rule. Un-analysed items remain unknown words, and the output of the analysis is almost perfect, which is an important condition for the rest of the process (i.e. transfer and generation).

6.3 Evaluation of the performance of the generation

Generation can also be evaluated according to two points of view: the correctness of the generated items, and the improvement brought by the solved words to the quality of the translated sentence.

To evaluate the first aspect, many procedures can be put in place. The correctness of constructed words could be evaluated by human judges, but this kind of approach would raise many questions and biases: people that are not expert of morphology would judge the correctness according to their degree of *acceptability* which varies between judges and is particularly sensitive when neologism is concerned. Questions of homogeneity in terms of knowledge of the domain and of the language are also raised.

Because of these difficulties, we prefer to centre the evaluation on the existence of the generated neologisms in a corpus. For neologisms, the most adequate corpus is the Internet, even if the use of such an uncontrolled resource requires some precautions (see (Fradin, Dal et al. 2007) for a complete debate on the use of web resources in morphology).

Concretely, we use the robot Golf (Thomas 2008) that sends each generated neologism automatically as a request on a search engine (here Google©) and reports the number of occurrences as captured by Google. This robot can be parameterized, for instance by selecting the appropriate language.

Because of the uncontrolled aspect of the resource, we distinguish three groups of reported frequencies: 0 occurrence, less than 5 occurrences and more than 5. The threshold of 5 helps to distinguish confirmed existence of neologism (> 5) from unstable appearances (< 5), that are closed to hapax phenomena.

The table below summarizes some results for some prefixed neologisms.

Prefix	tested forms	0 occ.	< 5 occ.	> 5 occ.
ri	391	8.2 %	5.6 %	86.2 %
anti	1120	8.6 %	19.9 %	71.5 %
de	114	2.6 %	3.5 %	93.9 %
super	951	28 %	30 %	42 %
pro	166	6.6 %	29.5 %	63.9 %
…				

Table 1 **:** Some evaluation results

Globally, most of the generated prefixed neologisms have been found in corpus, and most of the time with more than 5 occurrences. Unfound items are very useful, because they help to point out difficulties or miss-formalised processes. Most of the unfound neologisms were ill-analysed items in Italian. Others were due to misuses of hyphens in the generation. Indeed, in the program, we originally implemented the use of the hyphen in French following the established norm (i.e. a hyphen is required when the prefix ends with a vowel and the base starts with a vowel). But following this “norm”, some forms were not found in corpus (for example *antibraconnier* (Eng: *antipoacher*) reports 0 occurrence). When re-generated with a hyphen, it reports 63 occurrences. This last point shows that in neology, usage does not stick always to the norm.

The other problem raised by unknown words is that they decrease the quality of the translation of the entire sentence. To evaluate the impact of the translated unknown words on the translated sentence, we built up a test-suite of sentences, each of them containing one prefixed neologism (in bold in table 2). We then submitted the sentences to a commercial MT system (Systran©) and recorded the translation and counted the number of mistakes (FR1 in table 2 below). On a second step, we “feed” the lexicon of the translation system with the neologisms and their translation (generated by our prototype) and resubmit the same sentences to the system (FR2 in table 2).

For the 60 sentences of the test-suit (21 with an unknown verb, 19 with an unknown adjective and 20 with a unknown noun), we then counted the number of errors before and after the introduction of the neologisms in the lexicon, as shown below (errors are underlined).

IT	Le **defiscalizzazioni** logiche di 17 Euro sono previste	
FR1	Le defiscalizzazioni logiques de 17 Euro sont prévus	2
FR2	Les défiscalisations logiques de 17 Euro sont prévues	0

Table 2: Example of a tested sentence

For a global view of the evaluation, we classified in the table below the number of sentences according to the number of errors “removed” thanks to the resolution of the unknown word.

	0	**-1**	**-2**	**-3**
Nouns		10	8	2
Adjectives		18	1	
Verbs	2	14	3	2

Table 3: Reduction of the number of errors/sentence

Most of the improvements concern only a reduction of 1, i.e. only the unknown word has been solved. But it should be noticed that improvement is more impressive when the unknown words are nouns or verbs, probably because these categories influence much more items in the sentence in terms of agreement.

In two cases (involving verbs), errors are corrected because of the translation of the unknown words, but at the same time, two other errors are caused by it. This problem comes from the fact that adding new words in the lexicon of the system requires sometimes more information (such as valency) to provide a proper syntaxctic generation of the sentence.

6.4 Evaluation of feasibility and portability

The relatively good results obtained by the prototype are very encouraging. They mainly show that if the analysis step is performed correctly, the rest of the process can be done with not much further work. But at the end of such a feasibility study, it is useful to look objectively for the conditions that make such results possible.

The good quality of the result can be explained by the important preliminary work done (i) in the extension/specialisation of the lexicon, and (ii) in the setting up of the LFRs. The acquisition of the contrastive knowledge in a MT context is indeed the most essential issue in this kind of approach. The methodology we proposed here for setting these LFR proves to be useful for the linguist to acquire this specific type of knowledge.

Lexical morphology is often considered as not regular enough to be exploited in NLP. The evaluation performed in this study shows that it is not the case, especially in neologism. But in some cases, it is no use to ask for the impossible, and simply give up implementing the most inefficient rules.

We also show that the efficient analysis step is probably the main condition to make the whole system work. This step should be implemented with as much constraints as possible, to provide an output without errors. Such implementation requires proper evaluation of the impact of every constraint.

It should also be stated that such implementation (and especially knowledge acquisition) is time-consuming, and one can legitimately ask if machine-learning methods would do the job. The number of LFRs being relatively restrained in producing neologisms, we can say that the effort of manual formalisation is worthwhile for the benefits that should be valuable on the long term. Another aspect of the feasibility is closely related to questions of “interoperability”, because such implementation should be done within existing MT programs, and not independently as it was for this feasibility study.

Other questions of portability should also be considered. As we stated, we chose two morphologically related languages on purpose: they present less divergences to deal with and allow concentrating on the method. However, the proposed method (especially that contrastive knowledge acquisition) can clearly be ported to another pair of languages (at least inflexional languages). It should also be noticed that the same approach can be applied to other types of construction. We mainly think here of suffixation, but one can imagine to use LFRs with other elements of formation (like combining forms, that tend to be very “international”, and consequently the material for many neologisms). Moreover, the way the rules are formalised and the algorithm designed allow easy reversibility and modification.

7 Conclusion

This feasibility study presents the benefit of implementing lexical morphology principles in a MT system. It presents all the issues raised by formalization and implementation, and shows in a quantitative manner how those principles are useful to partly solve unknown words in machine translation.

From a broader perspective, we show the benefits of such implementation in a MT system, but also the method that should be used to formalise this special kind of information. We also emphasize the need for in-dept work of knowledge acquisition before actually building up the system, especially because contrastive morphological data are not as obvious as other linguistic dimensions.

Moreover, the evaluation step clearly states that the analysis module is the most important issue in dealing with lexical morphology in multilingual context.

The multilingual approach of morphology also paves the way for other researches, either in representation of word-formation or in exploitation of multilingual dimension in NLP systems.

References

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

Amiot, D. (2005) *Between compounding and derivation: elements of word formation corresponding to prepositions*. Morphology and its Demarcations. W. U. Dressler, R. Dieter and F. Rainer. Amsterdam, John Benjamins Publishing Company: 183-195.

Arnold, D., L. Balkan, R. L. Humphrey, S. Meijer and L. Sadler (1994) *Machine Translation. An Introductory Guide*. Manchester, NCC Blackwell.

Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston and M. Mazzoleni (2004) *Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian*. Proceedings of LREC 2004, Lisbon: 1771-1774.

Bouillon, P., S. Lehmann, S. Manzi and D. Petitpierre (1998) *Développement de lexiques à grande échelle*. Proceedings of Colloque des journées LTT de TUNIS, Tunis: 71-80.

Byrd, R. J. (1983) *Word Formation in Natural Language Processing Systems*. IJCAI: 704-706.

Byrd, R. J., J. L. Klavans, M. Aronoff and F. Anshen (1989) *Computer methods for morphological analysis* Proceedings of 24th annual meeting on Association for Computational Linguistics, New York, New York Association for Computational Linguistics: 120-127

Fradin, B., G. Dal, N. Grabar, F. Namer, S. Lignon, D. Tribout and P. Zweigenbaum (2007) *Remarques sur l'usage des corpus en morphologie*. Langages 167.

Gdaniec, C., E. Manandise and M. C. McCord (2001) *Derivational Morphology to the Rescue: How It Can Help Resolve Unfound Words in MT*. Proceedings of MT Summit VIII, Santiago Di Compostella: 127-131.

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

James, C. (1980) *Contrastive analysis*. Burnt Mill, Longman.

Maurel, D. (2004) *Les mots inconnus sont-ils des noms propres?* Proceedings of JADT 2004, Louvain-la-Neuve

Montermini, F. (2002) *Le système préfixal en italien contemporain*, Université de Paris X-Nanterre, Università degli Studi di Bologna**:** 355.

Namer, F. (2005) *La morphologie constructionnelle du français et les propriétés sémantiques du lexique: traitement automatique et modélisation*. UMR 7118 ATILF. Nancy, Université de Nancy 2.

Porter, M. (1980) *An algorithm for suffix stripping*. Program 14: 130-137.

Ren, X. and F. Perrault (1992) *The Typology of Unknown Words: An experimental Study of Two Corpora*. Proceedings of Coling 92, Nantes: 408-414.

Thomas, C. (2008) "Google Online Lexical Frequencies User Manual (Version 0.9.0)." Retrieved 04.02.2008, from http://www.craigthomas.ca/docs/golf-0.9.0-manual.pdf.

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