An Empirical Study for a Machine Aided Translation of French Prepositions 'à', 'de' and 'en' into English

Violaine Prince

*Montpellier University and LIRMM-CNRS

Abstract

This paper presents a study about ambiguous French prepositions, stressing out their roles as dependencies introducers, in order to derive some translation heuristics into English, based on a French-English set of parallel texts. These heuristics are formulated out of statistical observations and use some up-to-date results in Machine Translation (MT). Their originality mostly relies upon the importance given to syntax and dependency relations, along with lexicons, the latter being well browsed by the present literature in the domain. An experiment has been run on corpora in both languages, using a dependency parser in the source language, and results looked to be encouraging for a "step by step approach" for MT improvement.

1. Introduction

Researchers in in Computational Linguistics (CL) and Natural Language Processing (NLP) have been working on the role of prepositions and the issues they raise for an automated processing of language (Litkowski, 2005). When addressing the pair English-French, the impact of prepositions is a bit underestimated in the literature, when compared to the overwhelming importance given to nouns, noun phrases and adjectival phrases in lexical semantics. Electronic resources such as WordNet still assign to prepositions a secondary role, and mostly restricted to syntax, within multiword expressions. In Machine Translation (MT), a huge project such as ARTFL (American and French Research on the Treasury of the French Language), led by the University of Chicago and the French ATILF association for Machine Translation, provides an electronic bilingual resource of $75,000$ entries. Though, this resource seems to discard prepositional noun phrases acting as technical (and sometimes usual) multiword expressions. Although ontologies and web semantics have shed light on multiword expressions, several of which containing prepositions, as the lexical impersonation of some of their concepts, it seems that the real importance of prepositions is still left aside, whereas their role is quite crucial in the semantic interpretation of the fragment.

In this pair of languages, the preposition gives the direction in interpretation in a prepositional noun phrase. For instance *a set of pairs* greatly differs in meaning from *a pair of sets*. The non commutativity of natural language is here typically demonstrated by the existence and/or position of prepositions (Yeh and Vilain, 1998).

Second, the preposition semantics greatly modify the verb meaning in a phrasal verb, especially in English. Whole sets of new verbs are created by combining verbs with a strong polysemous potential such as *do*, *put*, *make*, *take*, *get*, etc. with prepositions (B. Trawinski and Soehn, 2006).

Third, prepositions are not univocal terms. They behave like nouns, adjectives and verbs by being as ambiguous as the latter (J.Taylor, 1993). For instance, if the preposition *at* seems to indicate a projection from a source towards a target in phrasal verbs such as *to look at*, a sense it shares with the preposition *to*, *at* seems a locative tag, with stability and immobility characteristics in *'he is at home'*, sharing the latter with the preposition *in* (Japkowicz and Wiebe, 1991). So characteristics such as polysemy and synonymy, common in nouns and verbs, also apply to prepositions.

The issue we try to address in this paper is the role and transformation of prepositions in a machine translation (MT) application in a restricted environment. The main items that we need to tackle are the following :

- Prepositions are often ambiguous in the source language (SL), but also their candidate equivalencies are ambiguous in the target language (TL). How can a task like MT disambiguate and choose the appropriate candidate?
- What is the syntactic role of the preposition? Does it introduce a particular complement? If yes, what is the influence of its semantics in SL on its representation in TL?
- Some prepositions belong to idiomatic multiword expressions and do not need to be examined as such. The recognition of their ineffectiveness as syntactic agents belongs to the existence (or the improvement) of existing translation memories and databases.

All these questions are naturally very broad in their scope. Since languages greatly vary in their use of linguistic dimensions, and in spite of an asymptotic aim at generalization, we have restricted their impact on the pair we address and all results will be considered as locally grounded in our English-French environment. They might possibly have an echo in other pairs of languages, and asserting the generality would be another very interesting issue that we cannot undertake in this paper.

2. Machine Translation and Prepositions : Is There a Most Appropriate Approach ?

The ground hypothesis is that MT, a very large field with several paradigms and software that contribute to its evolution, should be enhanced with tools able to detect as automatically as possible the nature of the dependency introduced by the preposition in SL, as a requirement for discriminating the proper equivalent in TL, provided that the preposition is not an element of an idiomatic multiword expression. Idiomatic expressions, as multiword colocations, can be stored in extensive lexicons, or translation memories. Their enhancement has been long studied in MT, and any survey, however extensive, will not capture all the achievements in this domain. A few works are still more

dedicated to prepositions (e.g., (Wehrli, 1998), (A. Villavicencio and Waldron, 2004)) or tend to stress them out (e.g, The Ultralingua dictionary at htttp://www.ultralingua.net/ provides human readable cross-lingual idiomatic expressions containing prepositions. The problem is that is not machine operable as such). Others (e.g. (V. Seretan, 2003)) have focused on the requirements for candidate idiomatic expressions in order to enhance existing resources. Several works have thoroughly contributed to the domain, but those we cite here have been the most inspiring to us, or the closest to our specific issue.

If the fragment to be translated is not idiomatic, then two approaches might compete:

- One that would browse aligned corpora and automatically learn translations and storing them in translation memories thus 'lexicalising' as much as possible the translation process. This is the actual main trend in MT. It needs important resources but few manpower.

Another that would undergo the same browsing, but would try to semi-automatically extract patterns to be rewritten into transformation rules. A more costly approach in terms of human effort, but since prepositions seem to deal with hidden construction aspects (e.g. syntax, dependencies) and less with apparent lexical data, it seems that a more rule based approach would not be totally out of scope, in such a case.

We have chosen to explore the second (and less used) path for which we have particular opportunities. If statistical MT was all successful in properly translating prepositions in the selected pair of languages , the issue would have dropped. However, the rationale is not "to beat" statistical MT. but to contemplate a hybrid method (both statistical and rule-based) seen as a task-properties oriented approach (if the local syntactic properties of prepositions provide a true added-value to their translation then a local rule-based tool within a statistical overall frame could be envisaged). Our research team has access to a dependency parser for French (Chauché, 2005), our source language, and to a French to English translation prototype that would implement, as a proof of concepts, some ideas about *transformation rules* as a formal rewriting of regular translation patterns. This prototype is compatible with the parser. But in order to do so, we needed to produce a study of prepositions behavior in both SL and TL, and this is what this paper is about. As computational linguists, we tried the most automatic approach, by parsing an SL corpus for which we had a TL aligned equivalent, we tagged parsed chunks with dependency tags, and studied the regularity (or not) of the prepositions translation according to their role.

3. Looking for Translation Patterns: An Experimental Framework

For a POC (Proof of Concepts) result, we have restricted the study to three prepositions widely used, which are *à*, *de* and *en*. The three could be seen as locative ((Japkowicz and Wiebe, 1991)) but are highly ambiguous, and might present a distributional similarity close to that described in (Baldwin, 2006). We have a pair of aligned French-English corpora of about $54,000$ words (in French) about stock exchange and economics, extracted from companies reports. The present corpus is specialized, "clean" and its quality is sufficient to ensure a reliable study. Moreover, it is naturally rich in prepositions. It is a good candidate "to begin with". We also used bilingual economics dictionaries [1] and an access to French and English Wordnet. First subsection states the issue, the second gives the general methods we have tried to follow and the third provides the obtained results.

3.1. A brief Overview of The Issue

This section describes the characteristics of the studied prepositions and their linguistic properties and the subsequent issues raised for computational representation. A particular light will be shed on the role of these prepositions as dependency triggers. As a consequence, we will address the particular problem of modeling translating actions for these prepositions.

3.1.1. Source Language Morphological Variation

Table 1 refer to lexical variability of prepositions '*de*' and ``a' in SL, whereas '*en*' is invariant. The two variants correspond to the contraction of a determinant (i.e. *le*, *les*) and the preposition and plays both the role of a preposition and a determinant. Some linguists consider only the determinant tag, some others insist on the dual role invoking the **substitution theory** (i.e, if one substitutes the singular feminine form of the determinant to the contracted token, it expands into two tokens, the preposition + the determinant. Thus, theoretically, *aux* should expand into *à les* , *du* should expand into *de le*, etc.). With substitution, one has to read "tag 1" and "tag 2" in table 2 not as mutually excluded tags, but as conjugated tags. The ambiguity here is not to be solved but is intrinsic to the contracted form, whereas with *en*, it is a classical mutually exclusive ambiguity.

Canonic Form	à	de
First Variant (Sing)	au	du
Second Variant (Plur.)	aux	des
Contracted	-	d'

Table 1: Different lexical forms representing the preposition

Token	Tag 1	Tag 2
du, des	determinant	preposition
au, aux	determinant	preposition
en	pronoun	preposition

Table 2: Different Part-of-Speech tags associated to lexical forms

[1] URL :http://www.e-anglais.com/ressources/glossary.html, developed by Kevin Halion. Mostly human readable, but machine operability was easy.

3.1.2. Dependency Roles

The different dependency roles of these prepositions are the following:
- in a noun phrase introducing a *noun complement* **(NC)**
- in a verb phrase introducing an *indirect noun phrase object complement* for verbs either transitive of intransitive **(OC)**
- in a verb phrase introducing an *infinitive proposition acting as an object complement* **(OCP)**
- in a verb phrase introducing *a noun phrase location complement* **(LC)**

Table 3 gives a few examples of translations with the three prepositions.

3.1.3. Multiple Equivalencies in Target Language

The ambiguity of the three studied prepositions is revealed and enhanced by the multiplicity of their equivalencies in the target language. Table 3 has already hinted at the fact that more than one equivalent is available for each of the prepositions. Table 4 shows what the bilingual dictionary we use suggests.

3.2. Modeling Prepositions Translation

One of the main items is the following question: Are prepositions part of a multiword expression? Do they already belong to a bilingual dictionary? If not, could they be possible candidates for domain-oriented translation lexicon?

The second question rises if prepositions are not part of multiword expressions. Some of the examples in table 3 show that translation sometimes deletes them. For instance, the NC (noun complement) dependency role seems to favor deletion of the preposition in TL (cf table 3). Could it be a transformation rule at the dependency level? Could it be flattened back at the pure component level in which the French Noun Phrase pattern *N1 Prep N2* (where *N1* and *N2* are nouns) , respectively *NP1 Prep NP2* (where *NP1* and *NP2* are noun phrases) would be translated in English into *TN2 ('s) TN1*, respectively *TNP2 TNP1* (where *Tx* is the translation of *x*)?

Prep.	Translations
à	at, to, in, with, by, upon, about
de	of, out of, off, from,with, by, about
en	in, into, to, of, thereof, at, during

Table 3: Equivalencies for each preposition

Another way to look at it is to consider the impact of dependency role on the translation. Does it reduce the translation ambiguity for a given preposition? If in the NC role, prepositions are deleted, in the LC (locative complement) role, is *à* always translated by *to*, *de* by *from*, and *en* by *by*? An experiment is interesting to conduct on such an issue.

Object complements (both OC and OCP roles) are more difficult to stress down. Here, a more thorough empirical study might lead to a better view of the subject.

3.2.1. Experimental Setup

We ran an experiment by parsing the SL corpus (54,000 words) with our morphosyntactic parser, which provides an deep syntactic analysis (with a syntactic tree) and assigns dependency roles to subtrees. The experiment was divided into three parts. It needed to answer the 'lexical or not lexical' question. Is the preposition an element of a multiword expression belonging or not to a lexical resource? For that we first checked that the parser was able to recognize lexical prepositional noun phrases. A second step aimed at tracking domain oriented multiword expressions by retrieving the most frequent subtrees containing any of our studied prepositions, in order to see if we can enhance a translation lexicon. The third investigated the remaining occurrences, focusing on dependency tags whenever they appeared.

3.3. Experiment Results

3.3.1. Checking the Parser

In our system, if the morphosyntactic parser recognizes the candidate as an idiomatic expression it transforms its pattern by adding '_%20' tags in the lexical string to replace the blanks. For instance, the idiomatic 'pomme de terre' meaning *potato* is transformed into 'pomme_%20de_%20terre'. The bilingual lexicon must contain it as a single entry. Other adjectival and verbal locutions exist in our dictionary, such as 'Beaucoup_%20de' (many), 'A_%20partir_%20de' (from... on) 'en_%20dépit_"%20de' (in spite of, despite). We created the machine readable bilingual lexicon out of our bilingual resources in such a format. We ran the parser on the SL corpus and obtained the results in table 5 according to the type of multiword expression. Two usual measures have been used: *recall* and *precision*. Recall is calculated as following: $\rho = \frac{nc}{nc+nf}$,
Where nc stands for the number of candidates correctly extracted and nf the number of candidates forgotten. Precision is calculated as: $\pi = \frac{nc}{np}$,
where np is the number of candidates extracted by the parser. Table 5 showing that idiomatic verbal locutions being less recognized by the parser, the latter was fed with this information and its abilities were thus enhanced.

3.3.2. Extracting the Most Frequent Noun Phrases with Prepositions

The second step was quite important in the sense that we needed to isolate sentence fragments containing prepositions that had proper syntactic roles, that were properly constituted and to which dependencies roles have been assigned. This meant that the corpus needed to be totally parsed.

The parser performances are not perfect: if POS tagging is over 98% in precision and recall ((Chauché, 2005)), syntactic analysis and dependency assignment, although quite successful compared to the on-going state-of-the-art, are closer to a 70% value. We had to run a tedious checking on the results and then, for well parsed subtrees, we re-run an automatic counting on the corpus.

As a first approach, we were interested into enhancing our bilingual lexicon with domain oriented expressions, and

Role	Example in SL	Translation
NC	*médecin **de** famille*	family doctor
	*moulin **à** café*	Coffeemill
	*pot **en** terre cuite*	terracotta pot
OC	*Je pense **à** lui*	I think **of** him
	*Je parle **de** lui*	I am talking **about** him
	Je pense en Franais	I think **in** French
OCP	*Je viens **de** manger*	I have just eaten
	*Je refuse **de** parler*	I refuse **to** talk
	*Il parle **en** dormant*	He talks **in** his sleep
LC	*Je pars **en** avion*	I am going **by** plane
	*Je pars **à** Londres*	I am going **to** London
	*Je viens **de** New York*	I am coming **from** New York

Table 4: Dependency Roles Examples

Type	Typical Example	Recall	Precision
Idiomatic NP	Hôtel de ville (Town Hall)	0.88	0.92
Idiomatic Nominal Locutions	En raison de (because of)	0.75	0.87
Idiomatic Verbal Locutions	A partir de (from ... on)	0.56	0.52
Idiomatic Adjectival Locutions	A peu près (circa)	0.90	0.88

Table 5: Extracted Candidates For Parser Checking

most of these are known to be noun phrases, so we extracted the 20 most frequent prepositional noun phrases. Since we had a TL corpus that was a translation of our SL corpus, we associated the TL equivalent and obtained the results given in table 6. The first observation was that the deletion of the preposition in the noun complement dependency role was quite regular, except in item 19 of table 6, so it was not a negligible clue for our future transformation rules. Second, it seems that the pattern *N1 Prep N2* in French is not always transformed into *TN2 TN1*. For instance, item 8 in table 6 gives birth to four English nouns, whereas items 12 pr 14 drop down to only one. Moreover,TN2 (respectively TN1) is not always the translation of French N2 (respectively French N1). 13 out of 20 expressions in table 6 do not follow such a pattern. So, in our opinion, *Domain Idiomatic Noun Phrases* would be those prepositional noun phrases in which nouns in TL are not translations of nouns in SL and should be candidate entries to the bilingual lexicon. Thus, we incorporated items 5,6,7,8,9,10,12,13,14,15,17,18,19 to our lexicon. Other noun phrases were extracted, and we noticed that there was a distribution, for the pattern *N1 de N2* (value "de" for the preposition) between *TN2 TN1* and *TN1 of TN2*, the latter being a word to word translation. These noun phrases were not specific to the domain.

3.3.3. Comparing Translations of Prepositions according to their Dependency Roles

If the NC role seems to lead to preposition deletion in quite an important number of cases and could be used as a rule if the expression is not idiomatic or domain specific, and if the preposition is not *de*, we tried to investigate other roles and the variation in translation within a given role. From the parsed corpus, we extracted all subtrees (chunks) in which prepositions *à*, *de* and *en* had one of the other dependency roles (within verb phrases). It is to be noticed that not all sentences were completely parsed, but all of them had at least a partial parsing (local subtrees formed). We had, at that point, two solutions: Either we checked all occurrences of those prepositions in a window of words (with an n-gram approach) on the aligned corpora (the SL and its translation), then assigned by hand dependency roles to our prepositional phrases, or we had to rely on the parsed SL corpus with its shortcomings. We chose to examine the well-formed verbal subtrees of the parsed SL corpus, containing our prepositions. Out of 1305 verbal phrases containing any of these prepositions, 962 were correctly parsed and the subtrees having the proper information in terms of dependency role tag [2]. Some of the reasons for such a recall value (i.e 0.73) were issued of a bad recall and precision on verbal locutions (cf table 5) or adverbial ones. The comparison of translations was made semi-automatically (with sentence numbers as usual in aligned corpora). The results are given in table 7.

What is quite interesting is the regularity of translations in the case of the locative complement (LC). *à* is translated as much by *at* as by *to*. The distribution has a semantic grounding: *at* is more static, pointing at the present place, whereas *to* is projective and points at the place to reach. In our corpus *de* is always translated by *from*, and *en* by *by*. Thus, the relationship with the semantic aspects is quite obvious: *en* designates the mean with which the movement is performed, and *de* the source. The only ambiguity is about the location, either present or future. The

[2]The parser developer has created a semi-automatic counting tool which allows him to check the parser abilities in terms of recall and precision.

Rank	Contents	Translation	Frequency
1	Economie de marché	market economy	45
2	Taux d'intérêt	Interest rate	42
3	Taux de croissance	growth rate	41
4	Taux d'inflation	inflation rate	39
5	Chiffre d'affaires	Sales turnover	39
6	Marché à terme	Futures exchange	37
7	Indice de référence	Benchmark	39
8	Indice de satisfaction	Customer Satisfaction Value Index	32
9	Consommation des ménages	Private Consumption	28
10	Marchés des devises	Exchange market	25
11	Augmentation de capital	Capital growth	23
12	Option d'achat	Call	19
13	jour de liquidation	Winding-up date	12
14	jour de valeur	Overnight	12
15	Obligations à haut rendement	Junk Bonds	7
16	Frais de souscription	Subscription charges	5
17	Frais de rachat	back-end load	5
18	Ecart de suivi	tracking error	4
19	Teneur de marché	mark to market	4
20	Facturation d'entreprise	Enterprise billing	3

Table 6: Most Frequent Prepositional Noun Phrases in the corpus (NC dependency role)

Role	à	de	en
OC	at $(22,8\%)$, to $(20,2\%)$, - ($18,1\%$), of$(16,2\%)$, with $(11,3\%)$, by $(6,6\%)$, about $(4,8\%)$	of $(37,2\%)$, -$(33,5\%)$ some $(17,9\%)$ by $(11,4\%)$	by $(40,3\%)$, in $(24,5\%)$ with $(10,4\%)$, - $(9,8\%)$ about$(9,5\%)$, into (5%)
OCP	to $(71,2\%)$, - ($28,8\%$)	- $(85,3\%)$, to (10%) about $(3,1\%)$, of $(1,6\%)$	in $(45,1\%)$ during (32%) while $(20,3\%)$, into $(3,4\%)$
LC	at ($50,9\%$) to $(49,1\%)$	from	by

Table 7: Prepositions translations (through alignment) and their percentages in the well parsed verbal phrases

OCP role, or the verb or proposition acting as an object complement, is more ambiguous for all prepositions, with a deletion case (represented by '-' in table 7) quite present. Let us note that preposition *en* is quite often translated in the meaning of a duration (*while*, *during*), especially with a present progressive verbal form. The widest distribution is for the OC (object complement) role assigned to a noun phrase. Here, it seems that the POS tag of the preposition (ambiguous as shown in table 2) has a role to play. When the tag is that of a determinant (case of *au*, *aux*, *du*, *des* as shown in table 1), then *to*, respectively - (i.e, the deletion of the preposition), are quite dominant, as well as *some*, an unlikely translation for *de*. It is also true for *en* : as a preposition, it is widely translated by *by*, whereas as a pronoun, it differs according to the nature of the object it refers to.

Last, a very interesting track has been suggested to us: Considering a 'verb/preposition' pair as a multiword expression *per se* and studying its semantic properties and the dependency roles attached to it. For this, one has to discard the classical tree building of Chauche's great parser for French and possibly 're-work' the corpus looking for POS tags bi-grams and thus producing patterns and associative rules that could be statistically exploited. This is a quite enjoyable research perspective and a work to be undertaken as a possible alternative to our present framework.

4. Conclusion

This first semi-automatic study on a corpus seemed to show the emergence of a few regular patterns, strongly related to the dependency role in case of noun and locative complements, and better discriminated when the appropriate POS tag is included in the case of noun phrases object complements to verbal phrases. The case of a verbal complement seems to indicate a preferred translation (see OCP line in table 7). This tends to mean that preposition translation obey to rules, and are far from being pure custom based (a case which would have favored a pure statistical approach). It also asserts that semantics have an important position (J.Taylor, 1993): Locative complements hint a spatial semantics, the behavior of *en* as a temporal complement appears in both noun phrases and OCP role and is translated by *in* or *during*, whereas its behavior as an instrument case indicates the use of *by*.

There is a very clear frame that appears if a precious resource such as a semantic parser, able to assign casual roles

to chunks in SL, is available as a pre-processor for machine aided translation. In our case, semantic roles were assigned in almost 90% of the correctly parsed OC and OCP, and all LC roles were correctly assigned. This gave us a valuable setting to produce transformation rules for our prototype, in which they were implemented (unfortunately, we have no room here to detail all the further work that has been undertaken and its results).

As an answer to the question concerning a possible generalization to either other prepositions or to another pair of languages, a broader experiment is currently being set up in order to investigate the prepositions behavior: 1) Outside the original corpus, extended to other corpora (focusing on the generic capacity of our results in a specific domain), 2) Compared to results returned by statistical MT devices (stressing out, or not the possible added value of semantic patterns), 3) Including other prepositions that would have similar properties in the same pair of languages, 4) Extending the framework to at least another pair (and relying upon the state-of-the art in such a pair). This is a huge research program and we just wanted to reach a reasonable outcome on the first item. Thus we intend to pursue the experiment by running the modified prototype on another corpus, and examining its success rate in properly translating

5. References

A. Villavicencio, T. Baldwin and B. Waldron, 2004. A multilingual database of idioms. In *Proceedings of the 4th International Conference On Language Resources and Evaluation, LREC-2004*.

B. Trawinski, M. Sailer and J-P. Soehn, 2006. Combinatorial aspects of collocational prepositional phrases. In P.Saint-Dizier (ed.), *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*. Kluwer Academic press, pages 197–210.

Baldwin, T., 2006. Distributional similarity and preposition semantics. In P.Saint-Dizier (ed.), *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*. Kluwer Academic press, pages 197–210.

Chauché, Jacques, 2005. Un analyseur du français en constituents et dépendances. Research Report 2005-138, LIRMM-CNRS.

Japkowicz, N. and J. Wiebe, 1991. A system for translating locative prepositions from english into french.'. In ACL (ed.), *Proceedings of the 29th annual meeting of the Association for Computational Linguistics*.

J.Taylor, 1993. *Prepositions : patterns of polysemization and strategies of disambiguation*. Walter de Gruyter, pages 151–178.

Litkowski, K. C., 2005. The preposition project. In ACL (ed.), *Proceedings of the Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*.

V. Seretan, E. Wehrli, L. Nerima, 2003. Multi-word collocation extraction by syntactic composition of collocation bigrams. In *Proceedings of RANLP*.

Wehrli, E., 1998. The preposition project. In ACL (ed.), *Proceedings of COLING-ACL*.

Yeh, A. S. and M.B. Vilain, 1998. Some properties of preposition and subordinate conjunction attachments. In ACL (ed.), *Proceedings of COLING-ACL*.