********** **********

Evelyne Viegas
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003

E-mail: viegas@crl.nmsu.edu
Fax: (505) 646 6218
Tel: (505) 646 5757

SIG-IL Pre-workshop Draft - Aug 28,1996
Do not quote - Very Rough Draft!

In this draft I will mainly address point 2), advocating that it takes an IL Text Meaning Representation (informed with planning techniques) to solve mismatches and divergences among various natural languages; and parts of point 3), in particular the different ways we

experimented in MikroKosmos to scale up "static" knowledge sources, to provide coverage of Spanish and English.

I - Point 2) Solving Mismatches and Divergences through an IL: a Case Study

Statementl: --------- It takes more than mere word sense disambiguation in analysis and lexical selection in generation to solve mismatches and divergences: it takes an IL Text Meaning Representation (TMR) informed by planning techniques.

In the following we will provide some empirical evidence from cross-linguistic data. We will first look at "simpler" cases of mismatches (such as "put" versus "polovit'" and "postavit'" in Russian) and then we will concentrate on the "continuum" that seems to exist between some mismatches and divergences as in "bake" and "cook" versus "cuire [+/- au four]" where only planning techniques seem to be able to generate the right lexeme or expression.

The following is brief, sketchy, and still needs argumentation...

Our interest for solving mismatches and divergences using an IL TMR along with planning, comes from noticing that all former enterprises (as described in Lindop et Tsujii, 1991; Door, 1990; Heid, 1993; Kameyama, 1991, Nirenburg and Levin, 1993, etc.) whatever the approach (or MT paradigm) seem to fail in solving (i.e., recognise and generate) divergences and mismatches. In terms of divergences (roughly speaking: same meaning but different syntactic structure) the problem seems to be linked to the impossibility to get an exhaustive typology of all the different types of divergences (cf Vandooren, 1993); moreover some cases seem difficult to classify, such as "wooden floor" -> "plancher" in French, similar to the conflation cases of Talmy (Talmy, 1985); or "bake" -> "cuire [+/- au four]" (where "au four" cannot be considered as a syntactic ellipsis). The case of mismatches (roughly speaking: the grammar and the lexicon of the SL do not make some distinctions which are required by the grammar and the lexicon of the TL) is even more problematic, as there is not only need for contextual knowledge but also for extra-linguistic knowledge, as discussed in (Kameyama, 1991).

Looking at real data from corpora, it seems that there are more examples which lay (still unexplained) in the continuum between divergences and mismatches than examples which can be classified as belonging to one case (clear example of predicative divergence: "he limped up the stairs" -> "il monta les marches en boitant") or the other (clear example of semantic underspecification "pez, pescado" ->

"poisson").

A big confusion wrt mismatches seems to arise from a largely shared belief that a language SL which has less lexical units to which correspond more lexical units in the TL (such as for "fish" in English -> "pez" and "pescado" in Spanish; or for "put" -> "polovit'" and "postavit'" in Russian; or for "cuire" in French -> "bake" and "cook"; ...) is ambiguous from a monolingual perspective.

To correct this supposed ambiguity one can decide there are two entries in the English dictionary for "fish" fish-N1 and fish-N2 corresponding to pez-N1 and pescado-N1 respectively. I believe a native English or American speaker to be very surprised to learn that where he had conceptualised one natural kind FISH he should now conceptualise two: FISH-living creature and FISH-food, without being able to make the link between the two, that is recognising the fact that what makes a fish a potential food, is the possibility of applying some cooking event to it in order to eat it (cf. Briscoe and Copestake, and their "grinding rule").

It rather seems to me that the word "fish" becomes ambiguous in Spanish while remaining unambiguous in English; same thing with polovit'/postavit' and put; or "bake/cook" and "cuire".

Isn't it rather the result of deliberate underspecification (elsewhere called vagueness) in some languages where inferences are sometimes preferred over short-cuts or fully specified meaning. Let me exemplified this with the Russian examples. I will consider the lexeme "put" as unambiguous in English but will have to consider it as underspecified wrt Russian.

I will assume a knowledge-based approach semantics based, and a conceptual world or ontology where i have a concept labeled PUT, which contains the following relevant information:

PUT
AGENT: HUMAN
THEME: PHYSICAL-OBJECT
SOURCE: PLACE
DESTINATION: PLACE

The semantics for "put", "polovit'" and "postavit'" should minimally have the following information:

put(X,Y,Z)
    sem: PUT(X,Y,Z), AGENT(X), THEME(Y), DESTINATION(Z)

polovit'(X,Y,Z)
sem: PUT(X,Y,Z), AGENT(X), THEME(Y), DESTINATION(Z),
DIRECTIONALITY(Y,FLAT)

postavit'(X,Y,Z)
sem: PUT(X,Y/Z)/ AGENT(X), THEME(Y), DESTINATION(Z),
DIRECTIONALITY(Y,UPRIGHT)

Now let us assume the following concepts GLASS and PLATE in the ontology with their associated conceptual relevant information:

GLASS
ISA: ARTIFACT
DIRECTIONALITY: UPRIGHT
CONTAINS: LIQUID

PLATE
ISA: ARTIFACT
DIRECTIONALITY: FLAT
CONTAINS: FOOD

Relevant extracts of an IL TMR for the simplified English sentence (a) John put the glass on the table, should look like:

PUT
AGENT: John
THEME: GLASS
DESTINATION: TABLE

Translating the above sentence into Russian does require some processing as there are two entries ("polovit'" and "postavit'") which can lexicalise the concept PUT. However, "polovit'" requires its theme to have a DIRECTIONALITY FLAT, which is the case of the word glass mapped to GLASS. Therefore mismatch viewed as specialisation (cf Kameyama, 1991) of lexical units is clearly a generation problem, not an analysis one.

Now if we look at the examples for "cook" and "bake" which translate into "cuire [+/- au four]", then here we seem to be confronted to a "generalisation" problem (cf. Kameyama, 1991). Here too we claim that we are confronted with a generation problem and not an analysis one as there is no reason to consider "cuire" as ambiguous in French. Now, let us consider the data below to illustrate the point that it takes

an IL TMR to solve mismatches and divergences.

Let us look in this draft at some isolated sentences, for the sake of simplicity:

b) Cuis le pain -> Bake the bread
c) Cuis les pa^tes al'dente -> Cook the pasta (al'dente)
d) Cuis les pa^tes au four -> d1) Bake the pasta
-> d2) Cook the pasta in the oven
e) Cuire les pa^tes au gratin
pas plus de 20mns -> e1) Bake the pasta au gratin no longer than 30mns
-> e2) Cook the pasta au gratin no longer than 30mns
f) I prefer baked meals to meals cooked on the stove top -> Je preferre les plats au four aux plats (cuisines) sur le feu
g) Cuire le pain et les pa^tes -> bake the bread, then cook the pasta

I said that "cuire" was not ambiguous in French. What remains to be seen is whether or not we get two concepts BAKE and COOK to which maps "bake" and "cook" respectively, with "cuire" mapping to COOK; therefore, going from English to French would be a question of generalisation whereas going from French to English would be a question of specialisation, as mentioned by (Kameyama, 1991). The problem with this approach is that it seems difficult in example f), which is a case of generalisation, to avoid to generate "je preferre des plats cuis a' des plats cuis sur le feu" (i preferred cooked meals to cooked meals on the stove)! Moreover, if we now want to specify, we have to rely on the semantics of the noun which sometimes is ambiguous, such as in example e) where although there is a preference for generating el) rather than e2), it is still acceptable to have e2). Finally, example g) shows that generating a mismatch requires more than lexical selection, it does require a planning of the sentence, as the conjunction "et" in French might be interpreted as a temporal-succession in which case it is necessary to develop the ellipsis. Moreover, contextual constraints present in the TMR will help to eventually generate bake the pasta if in the linguistic context we are told that "pasta" is a reference for "lasagna". The point remains that it is impossible to "freeze" the meanings of "bake" and "cook" as equivalent to "cuire au four" and "cuire" respectively, this is why I advocate an IL TMR along with planning to solve cross-linguistic problems of this kind.

Information to be included in the knowledge sources:

COOK
  AGENT: HUMAN
  THEME: PHYSICAL-OBJECT
  INSTRUMENT: COOKING-EQUIPMENT
  LOCATION: PLACE

cook(X,Y)
        sem: COOK(X,Y), AGENT(X), THEME(Y)

bake(X,Y)
        sem: COOK(X,Y), AGENT(X), THEME(Y), INSTRUMENT(OVEN)

cuire(X,Y)
        sem: COOK(X,Y), AGENT(X), THEME(Y),
INSTRUMENT(COOKTNG-
EQUIPMENT)

(to be developed; compare with "i started cooking at 18" -> cuisiner)

II - 3) Scaling up the "static" knowledge sources

Statement 2: Scaling up static knowledge sources to perform coverage
---------- is doable within a contemplative view of the lexicon: we
          did it!

The most difficult task seems to get started, namely get the core lexicon. In Mikrokosmos we developed a computational semantic lexicon for Spanish; each entry containing in the semantic zone an "unsaturated piece of IL-TMR". A core lexicon of about 7000 entries (lexemes) have been acquired by hand, with the use of computational tools to accelerate acquisition (lextool interface for acquisition; corpora search; on-line dictionary search; ontology browser; ontology request...). Then, we extended the core lexicon using derivational morphology applied to verbs, reaching around 35,000 entries-lexemes (for which we can produce the POS, the syntax and the semantics).

The big advantage of using an IL representation to encode the meanings of words is that the analysis lexicon can be reversed or indexed on concepts; this allowed us to perform many "exercises" as varied as:

- use the "reversed" lexicon as a pivot lexicon in a multilingual generation environment, by lexicalising in different languages the

semantic zone. For instance, the conceptual frame:

INGEST
AGENT(X)
THEME(Y)
EDIBLE(Z)

coming from the Spanish verb "comer-V1", can be lexicalised as "manger" in French, "eat" in English, etc... it can also serve as the basis for the lexicalisations of "close synonyms" "avaler, ingurgiter", ..., in French. Note that parallel corpora could also be used to see how "comer" translates into other languages; however, there still will be a need for human checking, but this should be faster than developing another lexicon from scratch, as our experience showed.

- we can generate from the TMR the text in Spanish and then analyse the gaps between the original source text and the text generated, this could enhance a lot the issue of what to put and what to omit in the IL and also how good our lexicons are.

Statement 3: Before scaling up for coverage there is still many work
---------- to be investigated if we adopt an inquisitive view of the lexicon (how useful it is wrt a particular task).

In the previous statement, I claimed it is doable to get coverage in a fairly small amount of time (it took us about a year with 4 person/year to develop a Spanish lexicon of about 35,000 roots, from scratch).

Here I would like to defend the position that the advantage of using an IL TMR lays in the power it gives us to capture meanings across languages. From the point of view of the "static" knowledge sources, the trade-offs between the lexicon and the ontology, calls sometimes for procedures (such as specialisation or specification or planning) not yet fully understood; however, I do believe that IL has more than any other approach to give us to capture the meaning(s) of words. The question of scaling up for coverage is not particular to IL approaches it is a common problem faced by any symbolic approach, and as such I do not think we should spend too much time on it. I guess we should rather try to explain unsolved phenomena, recognising which procedures to use to solve them. From the lexicon point of view, work on "underspecification" might well be the way to reconcile the contemplative view with the inquisitive view of the lexicon.