HANDLING SYNTACTICAL AMBIGUITY IN MACHINE TRANSLATION

Vladimir Pericliev

Institute of Industrial Cybernetics and Robotics
Acad. G.Bontchev Str., bl.12
1113 Sofia, Bulgaria

ABSTRACT

The difficulties to be met with the resolution of syntactical ambiguity in MT can be at least partially overcome by means of preserving the syntactical ambiguity of the source language into the target language. An extensive study of the correspondences between the syntactically ambiguous structures in English and Bulgarian has provided a solid empirical basis in favor of such an approach. Similar results could be expected for other sufficiently related languages as well. The paper concentrates on the linguistic grounds for adopting the approach proposed.

1. INTRODUCTION

Syntactical ambiguity, as part of the ambiguity problem in general, is widely recognized as a major difficulty in MT. To solve this problem, the efforts of computational linguists have been mainly directed to the process of analysis: a unique analysis is searched (semantical and/or world knowledge information being basically employed to this end), and only having obtained such an analysis, it is proceeded to the process of synthesis. On this approach, in addition to the well known difficulties of general-linguistic and computational character, there are two principle embarrassments to be encountered. It makes us entirely incapable to process, first, sentences with "unresolvable syntactical ambiguity" (with respect to the disambiguation information stored), and, secondly, sentences which must be translated ambiguously (e.g. puns and the like).

In this paper, the burden of solution of the syntactical ambiguity problem is shifted from the domain of analysis to the domain of synthesis of sentences. Thus, instead of trying to resolve such ambiguities in the source language (SL), syntactically ambiguous sentences are synthesized in the target language (TL) which preserve their ambiguity, so that the user himself rather than the parser disambiguates the ambiguities in question.

This way of handling syntactical ambiguity may be viewed as an illustration of a more general approach, outlined earlier (Penchev and Pericliev 1982, Pericliev 1983, Penchev and Pericliev 1984), concerned also with other types of ambiguities in the SL translated by means of syntactical, and not only syntactical, ambiguity in the TL.

In this paper, we will concentrate on the linguistics grounds for adopting such a manner of handling of syntactical ambiguity in an English into Bulgarian translation system.

2. PHILOSOPHY

This approach may be viewed as an attempt to simulate the behavior of a man-translator who is linguistically very competent, but is quite unfamiliar with the domain he is translating his texts from. Such a man-translator will be able to say what words in the original and in the translated sentence go together under all of the syntactically admissible analyses; however, he will be, in general, unable to make a decision as to which of these parses "make sense". Our approach will be an obvious way out of this situation. And it is in fact not infrequently employed in the everyday practice of more "smart" translators.

We believe that the capacity of such translators to produce quite intelligible translations is a fact that can have a very direct bearing on at least some trends in MT. Resolving syntactical ambiguity, or, to put it more accurately, evading syntactical ambiguity in MT following a similar human-like strategy is only one instance of this.

There are two further points that should be made in connection with the approach discussed. We assume as more or less self-evident that:

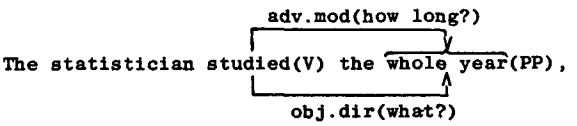
(i) MT should not be intended to explicate texts in the SL by means of texts in the TL as previous approaches imply, but should only translate them, no matter how ambiguous they might happen to be;

(ii) Since ambiguities almost always pass unnoticed in speech, the user will unconsciously disambiguate them (as in fact he would have done, had he read the text in the SL); this, in effect, will not diminish the quality of the translation in comparison with the original, at least insofar as ambiguity is concerned.

3. THE DESCRIPTION OF SYNTACTICAL AMBIGUITY IN ENGLISH AND BULGARIAN

The empirical basis of the approach is provided by an extensive study of syntactical ambiguity in English and Bulgarian (Pericliev 1983), accomplished within the framework of a version of dependency grammar using dependency arcs and bracketings. In this study, from a given list of configurations for each language, all logically-admissible ambiguous strings of three types in English and Bulgarian were calculated. The first type of syntactically ambiguous strings is of the form:

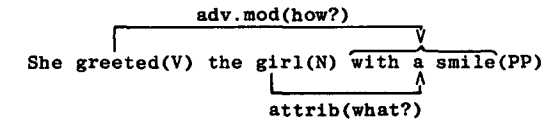
(1) A $\overset{1/2}{\text{---}}$>B, e.g.



where A, B, ... are complexes of word-classes, "--->" is a dependency arc, and 1, 2, ... are syntactical relations.

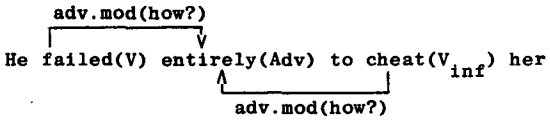
The second type is of the form:

(2) A $\overset{1}{\text{--}}$>B<$\overset{2}{\text{--}}$ C, e.g.



The third type is of the form:

(3) A $\overset{1}{\text{--}}$>B<$\overset{1}{\text{--}}$ C, e.g.



It was found, first, that almost all logically-admissible strings of the three types are actually realized in both languages (cf. the same result also for Russian in Jordanskaja (1967)). Secondly, and more important, there turned out to be a striking coincidence between the strings in English and Bulgarian; the latter was to be expected from the coincidence of configurations in both languages as well as from their sufficiently similar global syntactic organization.

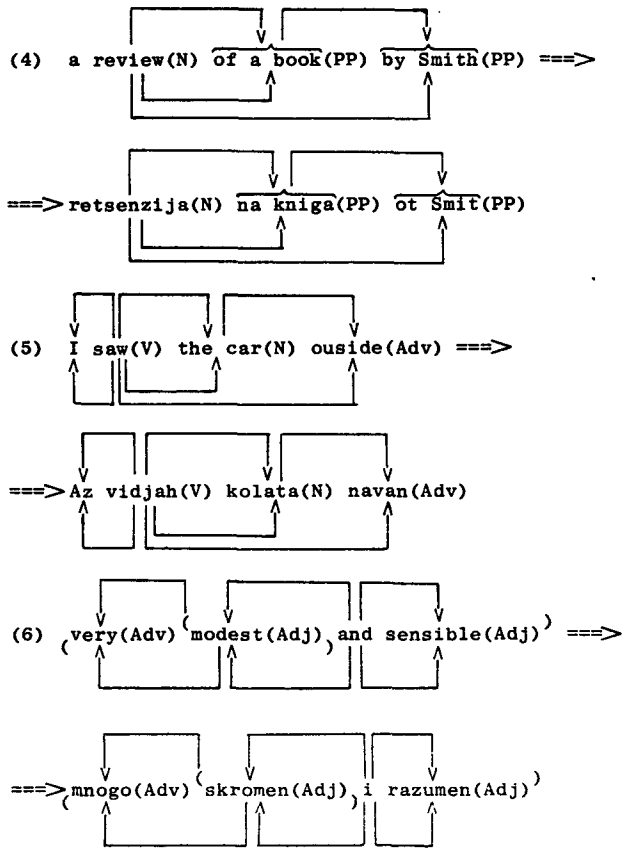
4. TRANSLATIONAL PROBLEMS

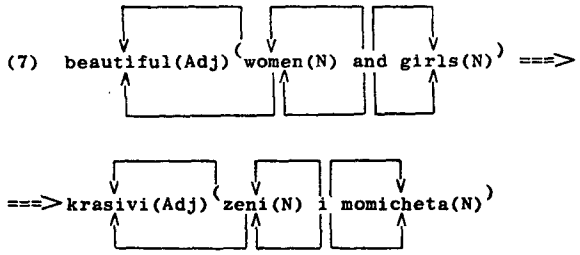
With a view to the aims of translation, it was convenient to distinguish two cases: Case A, in which to each syntactically ambiguous string in English corresponds a syntactically ambiguous string in Bulgarian, and Case B, in which to some English strings do not correspond any Bulgarian ones; Case A provides a possibility for literal English into Bulgarian translation, while there is no such possibility for sentences containing strings classed under Case B.

4.1. Case A: Literal Translation

English strings which can be literally translated into Bulgarian comprise, roughly speaking, the majority and the most common of strings to appear in real English texts. Informally, these strings can be included into several large groups of syntactically ambiguous constructions, such as constructions with "floating" word-classes (Adverbs, Prepositional Phrases, etc. acting as slaves either to one, or to another master-word), constructions with prepositional and post-positional adjuncts to conjoined groups, constructions with several conjoined members, constructions with symmetrical predicates, some elliptical constructions, etc.

Due to space limitations, a few English phrases with their literal translations will suffice as an illustration of Case A. (Further on, syntactical relations as labels of arcs will be omitted where superfluous in marking the ambiguity):



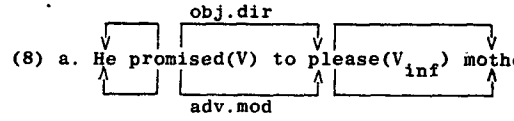


4.2. Case B: Non-Literal Translation

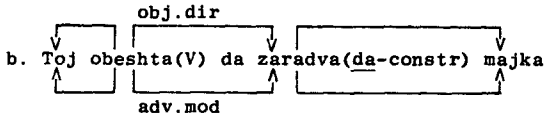
English strings which cannot be literally translated into Bulgarian are such strings which contain: (i) word-classes (V_{inf}, Gerund) not present in Bulgarian, and/or (ii) syntactical relations (e.g. "composite": language<--- theory, etc.) not present in Bulgarian, and/or (iii) other differences (in global syntactical organization, agreement, etc.).

It will be shown how certain English strings falling under this heading are related to Bulgarian strings preserving their ambiguity. A way to overcome difficulties with (ii) and (iii) is exemplified on a very common (complex) string, viz. Adj/N/Prt+N/N's+N (e.g. stylish gentlemen's suits).

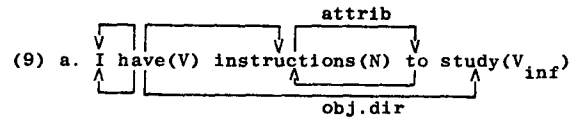
As an illustration, here we confine to problems to be met with (i), and, more concretely, to such English strings containing V_{inf}. These strings are mapped onto Bulgarian strings containing da-construction or a verbal noun (V_{inf} generally being translated either way). E.g. the V_{inf} in



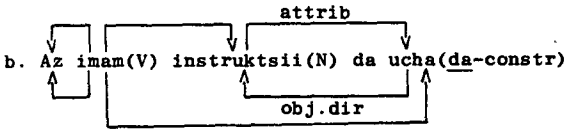
(promised what or why?) is rendered by a da-construction in agreement with the subject, preserving the ambiguity:

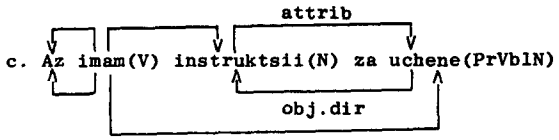


In the string



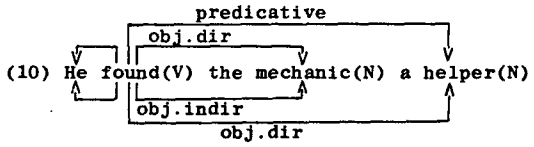
(what instructions or I have to study what?) V_{inf} can be rendered alternatively by a da-construction or by a prepositional verbal noun:





Yet in other strings, e.g. The chicken(N) is ready(Adj) to eat(V_{inf}) (the chicken eats or is eaten?), in order to preserve the ambiguity the infinitive should be rendered by a prepositional verbal noun: Pileto(N) e gotovo(Adj) za jadene (PrVblN), rather than with the finite da-construction, since in the latter case we would obtain two unambiguous translations: Pileto e gotovo da jade (the chicken eats) or Pileto e gotovo da se jade (the chicken is eaten), and so on.

For some English strings no syntactically ambiguous Bulgarian strings could be put into correspondence, so that a translation with our method proved to be an impossibility. E.g.



(either the mechanic or someone else is the helper) is such a sentence due to the impossibility in Bulgarian for two non-prepositional objects, a direct and an indirect one, to appear in a sentence.

4.3. Multiple Syntactical Ambiguity

Many very frequently encountered cases of multiple syntactical ambiguity can also be handled successfully within this approach. E.g. a phrase like Cybernetical devices and systems for automatic control and diagnosis in biomedicine with more than 30 possible parsings is amenable to literal translation into Bulgarian.

4.4. Semantically Irrelevant Syntactical Ambiguity

Disambiguating syntactical ambiguity is an important task in MT only because different meanings are usually associated with the different syntactical descriptions. This, however, is not always the case. There are some constructions in English the syntactical ambiguity of which cannot lead to multiple understanding. E.g. in sentences of the form A is not B (He is not happy), in which the adverbial particle not is either a verbal negation (He isn't happy) or a non-verbal negation (He's not happy), the different syntactical trees will be interpreted semantically as synonymous: 'A is not B' <===> 'A is not-B'.

We should not worry about finding Bulgarian syntactically ambiguous correspondences for such English constructions. We can choose arbitrarily one analysis, since either of the syntactical descriptions will provide correct information for our translational purposes. Indeed, the construction above has no ambiguous Bulgarian correspondence: in Bulgarian the negating particle combines either with the verb (then it is written as a separate word) or with the adjective (in which case it is prefixed to it). Either construction, however, will yield a correct translation: Toj ne e radosten or Toj e neradosten.

4.5. A Lexical Problem

Certain difficulties may arise, having managed to map English syntactically ambiguous strings onto ambiguous Bulgarian ones. These difficulties are due to the different behavior of certain English lexemes in comparison to their Bulgarian equivalents. This behavior is displayed in the phenomenon we call "intralingual lexical-resolution of syntactical ambiguity" (the substitution of lexemes in the SL with their translational equivalents from the TL results in the resolution of the syntactical ambiguity).

For instance, in spite of the existence of ambiguous strings in both languages of the form $\text{Verb}_{tr/itr} \xrightarrow{obj/subj} \text{Noun}$, with some particular lexemes (e.g. $\text{shoot}_{tr/itr} \Longrightarrow \text{zastreljam}_{tr}$ or streljam_{itr}), in which to one English lexeme correspond two in Bulgarian (one only transitive, and the other only intransitive), the ambiguity in the translation will be lost. This situation explains why it seems impossible to translate ambiguously into Bulgarian examples containing verbs of the type given, or verbal nouns formed from such verbs, as the case is in The shooting of the hunters. This problem, however, could be generally tackled in the translation into Bulgarian, since it is a language usually providing a series of forms for a verb: transitive, intransitive, and transitive/intransitive, which are more or less synonymous (for more details, cf. Penchev and Pericliev (1984)).

5. CONCLUDING REMARKS

To conclude, some syntactically ambiguous strings in English can have literal, others non-literal, and still others do not have any correspondences in Bulgarian. In summary, from a total number of approximately 200 simple strings treated in English more than 3/4 can, and only 1/4 cannot, be literally translated; about half of the latter strings can be put into correspondence with syntactically ambiguous strings in Bulgarian preserving their ambiguity. This gives quite a strong support to the usefulness of our approach in an English into Bulgarian translation system.

Several advantages of this way of handling of syntactical ambiguity can be mentioned.

First, in the processing of the majority of syntactically ambiguous sentences within an English into Bulgarian translation system it dispenses with semantical and world knowledge information at the very low cost of studying the ambiguity correspondences in both languages. It could be expected that investigations along this line will prove to be fruitful for other pairs of languages as well.

Secondly, whenever this way of handling syntactical ambiguity is applicable, the impossibility of previous approaches to translate sentences with unresolvable ambiguity, or such with verbal jokes and the like, turns out to be an easily attainable task.

Thirdly, the approach seems to have a very natural extension to another principal difficulty in MT, viz. coreference (cf. the three-ways ambiguity of Jim hit John and then he (Jim, John or neither?) went away and the same ambiguity of toj (=he) in its literal translation into Bulgarian: Djim udari Djon i togava toj(?) si otide).

And, finally, there is yet another reason for adopting the approach discussed here. Even if we choose to go another way and (somehow) disambiguate sentences in the SL, almost certainly their translational equivalents will be again syntactically ambiguous, and quite probably preserve the very ambiguity we tried to resolve. In this sense, for the purposes of MT (or other man-oriented applications of CL) we need not waste our efforts to disambiguate e.g. sentences like John hit the dog with the long bat or John hit the dog with the long wool, since, even if we have done that, the correct Bulgarian translations of both these sentences are syntactically ambiguous in exactly the same way, the resolution of ambiguity thus proving to be an entirely superfluous operation (cf. Djon udari kucheto s dalgata palka and Djon udari kucheto s dalgata valna).

6. REFERENCES

Jordanskaja, L. 1967. Syntactical ambiguity in Russian (with respect to automatic analysis and synthesis). Scientific and Technical Information, Moscow, No.5, 1967. (in Russian).

Penchev, J. and V. Pericliev. 1982. On meaning in theoretical and computational semantics. In: COLING-82, Abstracts, Prague, 1982.

Penchev, J. and V. Pericliev. 1984. On meaning in theoretical and computational semantics. Bulgarian Language, Sofia, No.4, 1984. (in Bulgarian).

Pericliev, V. 1983. Syntactical Ambiguity in Bulgarian and in English. Ph.D. Dissertation, ms., Sofia, 1983. (in Bulgarian).