SOME PROBLEMS OF MACHINE TRANSLATION BETWEEN CLOSELY RELATED LANGUAGES

Alevtina BENOVA, Karel OLIVA and Jarmila PANEFOVA
Faculty of Mathematics and Physics
Charles University
Malostranské náměstí 25
CS-118 00 Praha 1 - Malá Strana
Czechoslovakia

Abstract:
We describe the linguistic background of a Czech-to-Russian MT system, stressing its features resulting from the close relatedness of the two languages, above all the possibility of a minimization of the transfer. Related linguistic problems are analysed within the MT project, as well as in the perspective of contrastive linguistics.

1. The system of Czech-to-Russian MT system called RUSLAN is conceived (similarly as all linguistically based MT systems) as a modular system consisting (in brief) of a source language parser, a transfer and a synthesis of the target language. The task is to translate texts from the domain of computers, in particular manuals of operating systems. Since in RUSLAN the source language is closely genetically related to the target one, some of the modules of the system could have been considerably simplified, not leaving out of consideration the theoretical linguistic framework on which the system is based (dependency and stratificational approach). The simplifications concern, first of all, the transfer phase, so that the system cannot be understood as including a complete transfer.

2. The effort towards a maximally effective procedure has also resulted in simplifications in the parser. This was made possible i.a. by the similarity of cases of syntactic ambiguity in the source and the target language. For example, with sequences of the type Verb Noun1 Noun2 ... Nounm, where each Noun stands for a nominal or a prepositional group, a free modifier, the surface order can generally be preserved, which fact makes unnecessary a detailed identification whether any of the Nouns modifies the Verb or one of the preceding Nouns. This can be illustrated by the output Russian sentence "Vo vremja svojej raboty programa mozhet potrebovat' takse pomoč.' sistemy pri obrabotke fajlov dannych." (Lit. "In course of its work program can need also help of system in processing of-files of-data."). Here, where the verb "potrebovat'" can be translated as "to need", the passive "system" is selected as the passive surface subject, expressed then by nominative. With each of the slots, the semantic features required or excluded for the filler of that slot are indicated. These features help to identify the fillers, especially in cases of ambiguity, e.g. in Czech "Výstupní zařízení nastaví řídicí na požadovanou hodnotu." (Lit. "Output device sets line-spacing at required value."). The verb "nastavit" has the following valency frame: Actor (nom/nom, +Human,+Device), Objective (ace/ace, +Con-cer,+Result-of-process,+Human), where + denotes semantic features such that at least one of them has to be present with the filler of the respective slot, "-" denotes semantic features excluded with the filler, and boldface denotes Czech-Russian morphological forms. In this way, the ambiguity of morphemic case with "řídicí" and "zařízení" (in both cases between nom and ace) can be solved on the basis of semantic features of the two nouns.

3. The choice of the Russian equivalents for Czech lexical units should reflect also...
structural differences between the two languages. These differences concern also syntactic patterns; at least the following cases should be distinguished:

a. Adj Adj Noun -&gt; Adj Noun
   ex.: data control command
   lit.: data control command

b. Noun -&gt; Adj Noun
   ex.: poštka -&gt; vyšištelnáma mašina
   lit.: computer -&gt; computing machine

c. Verb -&gt; Verb Noun
   ex.: ekompilovat -&gt; osuššoství komplikaci
   lit.: to compile -&gt; to carry out compilation

d. Noun -&gt; Noun Noun
   ex.: poštka -&gt; taška perenečení
   lit.: beginning -&gt; point of intersection

e. Adj Adj Noun -&gt; Noun Noun Adj Noun
   ex.: vyšší programovací jazyk
   lit.: higher programming language

Clearly, some types are easier to implement than the others, which depends on the complexity of the respective Czech and Russian constructions. For simplification of some cases of the type a., where the Russian equivalent includes a modifying noun in a fixed morphemic form, this is treated as an uninflected word, the syntactic relation of which is established already in the dictionary. Due to the closeness of the languages, a useful ingredient can be seen in the idea of a transducing dictionary proposed and elaborated in the English-to-Czech MT system (cf. Kirshner,82). The transducing dictionary, based on algorithmic handling of the regular productive international affixes (with exceptions listed in the main dictionary) and of the orthographic and similar differences, can be illustrated by the following:

a. with the suffixes -ř (more, "assembly"), -š (aggregate, "aggrega"); -nt (coefficient, "coefficient"), -rpm (rpm, "cubic volume"), and the lexical components of Greek or Latin origin, such as -graf, -úkop (radiographic, "radiograph"); the Russian equivalents differ at most in details.

b. with other suffixes of international use, the Russian equivalents correspond in a systematic way to the Czech ones, as with
   -leta/-lat, -ov/-ja, -imenn/-im,
   -rúf/-rúmy, -lokt/-lokti

c. to a certain degree also words of Slavic origin can be handled by a procedure based on correspondences with regular segment pairs such as k/g, u/r, Trč/Torc (where r stands for an obvious krč/kořtikj "short"), such pairs as hrud ("chest") vs. "gorod ("town"), where the lexical semantics differ, have to be listed in the lexicon.

d. whenever a word has not been identified in the main dictionary and cannot be treated by the procedures of the types a.,b.,c., at least transliteration and some of the elementary correspondences are carried out, so that if e.g. "přeplňen" ("overloading") or "dis-
keta" ("floppy disk") were not found in the dictionary, they would be transduced as "pe-
repolnenie" (correctly) and "disketa" (in
stead of "diskjisk"), respectively.

The procedure, and a set of similar fall-back rules for syntax, should ensure that the output be basically understandable.

4.4.1 One of the relevant differences between Czech and Russian syntax concerns sentences with the Czech first person plural corresponding to the Russian reflexive form; e.g. Czech "Algoritmus rozmeščení bloků popínajíme v šestu 6" vs. Russian "Algoritmus razmeščení bloků opíjvajúc se v rozadle 6" ("The algorithm of dislocation of blocks is described in Sect. 6"). Often a modal expression is present: "Ráby programu můžeme najít v knihovně" vs. Russian "Razvádží programu možno najít v bibliotece" ("The titles of the programs can be found in the library"). The linguistic rules underlying the practical solution of these problems can have the following form:

\[ \text{Noun}_{acc} \text{Verb}_{stpl} \rightarrow \text{Noun}_{nom} \text{Verb}_{refl} \]

\[ \text{Noun}_{acc} \text{Verb}_{modal}_{stpl} \text{Verb}_{inf} \rightarrow (\text{Noun}_{nom}) \text{ Modal Verb inf} \]

("Modal" stands here for such expressions as "možno" ("possible"), "nado" ("necessary"); parentheses (') denote the fact that the Objective is not always obligatory.

4.4.2 In some cases the ambiguity of a Czech sentence corresponds to a similar ambiguity in Russian. In other cases the ambiguity in the two languages is not in such accordance. This is illustrated by the following:

a. Czech: V létě proběhlo jednání o nové variante OS.
   Russian: Letom prošlo sovětisaté o nové variante OS.
   (In summer, the negotiations on the new vari-
   ant of OS took place.)

b. Czech: V létě proběhlo jednání o prázdních.
   Russian: Letom sovětisaté prošlo ve vremja kantrul.
   (In summer, the negotiations took place during vacations.)
The preposition "o" with locative in Czech is kept also in Russian or, with nouns having the feature Time, translated as "vo vremya" with genitive.

Differences in prepositional constructions are found also with the following pairs:

c. Czech:
Práce na programu pokračují i v tomto roce.
Raboty nad programmoy prodolžajuta i v tom rodu.
(Translation: The works on the program continue also this year.)
d. Czech:
Práce na fakultě pokračují i v tomto roce.
Raboty na fakułte prodolžajuta i v tom rodu.
(Translation: The works at the faculty continue also this year.)

These examples cannot be fully accounted for by means of lexical information, neither can they be included into the general scheme of syntactic rules. It is necessary to have a list of such differences.

4.3 In translating Czech subordinate clauses introduced by such conjunctions as "zda", "-li", "jestliže" ("if"), "když" ("when"), "dokud" ("until"), "podruhé" ("then"), some of which are ambiguous, the text can be treated as relatively homogenous. The functioning of a clause introduced by "zda" or "-li" as a subject can be identified on the basis of the valency of the verb in superordinated clause, where it is marked whether the verb may take a subordinated clause as its Actor or Objective. In the other cases, suitable or at least acceptable translations of the conjunctions are as follows: Czech "zda", "-li", "podruhé" as Russian "esli"; Czech "když", "dokud ne" as Russian "poka", "poka ne", Czech "když" as Russian "kogda".

It follows that while it is necessary to work to a certain degree with the underlying structure, in the majority of cases the equivalent can be chosen just in accordance with the conjunctions themselves.

4.4 The Czech verb "být" ("to be") has several Russian equivalents: the copula "byt'", verbs "est'", "javljat'sja", "nahodit'sja", "imeti'ja". The selection of the equivalent depends on the syntactic context: if the nominal predicate in Czech is in instrumental case, then a form of the verb "javljat'sja" is preferred; if a local adverbial is present, then the translation "nahodit'sja" is at place; otherwise the appropriate form of the copula is chosen. Of course, another point concerns the translation of "být" within idioms ("být' v porjádku", but "imeti'ya v ras-porjázhnom")

4.5 The surface behaviour of negation is not the same in Czech and in Russian: in Czech, even partial negation is often expressed as a prefix of the verb, which gives rise to an ambiguity absent in Russian, where this distinction is always transparent. Some of the examples from our texts are:

a. Czech:
To ani systém přesně neví.
Russian:
Nesetogo daša sistema to no ne znait.
(Translation: This even the system does not know exactly.)
b. Czech:
Tabulka není uložena na povném místě v paměti.
Russian:
Tablica pomalčajtaja ne na postojannom meste v pamjati.
(Translation: The matrix is not placed in a fixed position in the storage.)

4.6 We assume that the surface order is substantially the same in the two languages; the differences concern only such specific cases as, e.g., the positions of parts of the complex verb forms or those of certain pronouns and particles which have the character of elities in Czech, but usually follow the verb in Russian:

a. Czech:
... vypadal by tak, že by tabulka obsahovala
by...
Russian:
... vyigjadel by tak, što tablica podržala
by danny...
(... he would look as if the matrix con-
tained(COND.) data ...)
b. Czech:
Budeme se v operačních systémech snažit...
Russian:
V operacionnych sistemach budem sterhtaj...
(In the operating systems, we shall try ...)

The differences described in this section do not concern the structural order, and there is no danger that ambiguity might arise. The dislocation of function words and particles can be described by general rules.

4.7 In 4.1 through 4.6 we wanted to show what the problems of parsing are if the correspondences in the underlying structure, in surface syntax and in the surface order of morphemes are to be made use of, while the differences are solved; we also wanted to illustrate the narrowed, but nonetheless necessary role of transfer.

5. We wanted to point out that, on the one hand, the closeness of the two languages makes it relatively easy to find a strategy for an MT system, since the most complex problems of ambiguities might be partially avoided, although, on the other hand, comparative empirical research in the domains of lexicology and of syntax is necessary also for such a pair of languages. Results of such an approach may be useful in MT, and also in the context of a contrastive comparison of cognate languages.

References:
Kirschner Z.: On a Device in Dictionary Operations in Machine Translation, in proceedings of Coling '82, Prague
Panevová J.: FORMY A FUNKCE VE STAVBĚ ČESKÉ VĚTY, ACADENIA, PRAGUE, 1980