SOME PROBLEMS OF MACHINE TRANSLATION BETWEEN CLOSELY RELATED LANGUAGES

Alevtina BÉMOVÁ, karel OLÍVA and Jarmila PANEVOVÁ
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 closed relatedness of the two languages, above all the possibility of a minimization of the transfer. Related linguistic problems are analyzed 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 tranfer 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 Noun$_1$ Noun$_2$ ... Noun$_i$, where each Noun$_j$ stands for a nominal or a prepositional group serving as a free modifier, the surface order can generally be preserved, which fact makes unnecessary a detailed identification whether any of the Noun$_j$´s modifies the Verb or one of the preceding Nouns. This can be illustrated by the output Russian sentence "Vo vremja svoej raboty programma možet potrebovat' takže pomošč´ sistemy pri obrabotke fajlov dannych." (Lit. "In course of-its work program can need also help of-system in processing of-files of-data."), where the group "pri obrabotke ..." can be analyzed (in both languages) as modifying the verb "potrebovat'" or the nouns "pomošč´" or "sistemy". If the order of the nominal groups is preserved, the translation also preserves the structural ambiguity of the original. Also nominalizations can be translated independently of their underlying structure (e.g., "Indeksnoposledovateľnyje fajly neobchodimo do obrabotki preobrazovat'." - lit. "Index-sequential files have-to-be before processing transformed.", or "Programmy, napisannye na jazyke Assembler v ramkach predyduščej versii, neobchodimo snova translirovat'." - lit. "Programs written in language Assembler in framework-of preceding version have-to-be again compiled.").

Such an approach made it possible, at first, to minimize the transfer phase in the design of the project, and then, in the process of realization, the articulation of transfer operations into the parser and the synthesis, which may lead to an impression that RUSLAN works completely without transfer, i.e., as a direct binary MT system. In principle, it can be said that the minimization of the transfer reflects the empirical fact that the two languages have a lot of common features.

3. A great role is played in RUSLAN by the lexicon. The lexical entry contains maximum of information, which is then projected to the syntactic rules; only the most general behaviour of words is rendered purely by means of syntax.

The rules of choice of lexical equivalents include different types of information. Along with the data on parts of speech and morphemics, semantic features are listed, and (esp. with verbs) also the valency (subcategorization) frame; the valency slots are accompanied by information on their Czech morphemic form as well as that of the corresponding Russian items (as an example of their discrepancy might serve the pair "užívat něco(acc.)" vs. "poľzovat'sja čem(instr.)" - "to use stg."). Where pasivization is possible, it is indicated which of the slots (mostly, but not always expressed by accusative) 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 řádkování na požadovanou hodnotu." (lit. "Output device sets line-spacing at required value."), the verb "nastavit" ("set") has the following valency frame: Actor (**nom/nom**, +Human, +Device), Objective (**acc/acc**, +Concr, +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 bold-print denotes Czech/Russian morphological forms. In this way, the ambiguity of morphemic case with "řádkování" and "zařízení" (in both cases between nom and acc) can be solved on the basis of semantic features of the two nouns.

3.1 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 → Adj Noun
ex.: datový řídící příkaz
→ upravljajuščij operator
lit.: data control comand
→ control operator
b. Noun → Adj Noun
ex.: počítač → vyčislitel'naja mašina
lit.: computer → computing machine
c. Verb → Verb Noun
ex.: zkompilovat → osuščestvit' kompiljaciju
lit.: to compile → to carry out compilation
d. Noun → Noun Noun
ex.: počátek → točka peresečenija
lit.: beginning → point of-intersection
e. Adj Adj Noun → Noun Noun Adj Noun
ex.: vyšší programovací jazyk
→ jazyk programmirovanija vysšego urovnja
lit.: higher programming language
→ language of-programming of-higher level

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 d., 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.

3.2 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. Kirschner,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 -áž (montáž,"assembly"), -át (agregát,"aggregate"), -ent (koeficient, "coefficient"), -ura (kubatura,"cubic volume"), and the lexical components of Greek or Latin origin, such as -graf, -skop (kardiograf,"cardiograph",elektroskop,"electroscope"), 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 -ista/-ist, -ie/-ija, -ismus/-izm, -ární/-arnyj, -ický/-ičeskij
c. to a certain degree also words of Slavonic origin can be handled by a procedure based on correspondences with regular segment pairs such as h/g, ř/r, TraT/ToroT (where T stands for an occlusive: krátký/korotkij "short"); such pairs as "hrad" ("castle") vs. "gorod" ("town"), where the lexical semantics differs, 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. "přeplnění" ("overloading") or "disketa" ("floppy disc") were not found in the dictionary, they would be transduced as "perepolnenie" (correctly) and "disketa" (instead of "gibkij disk"), respectively.

This procedure, and a set of similar fail-soft rules for syntax, should ensure that the output be basically understandable.

4. The procedures of syntactic analysis and synthesis are based on lexical information, including the valency frames. Certain difficulties arise when filling the slots of obligatory adverbials (see Panevová,80) with which the forms of a given adverbial type are variable, e.g. "vrátit se kam" ("to return somewhere"): "napravo" ("to the right", adverb), "k problému" ("to the problem", preposition "k" + dative), "do bytu" ("into the flat", preposition "do" + accusative) etc. Such cases are handled by the parser together with free adverbials, only it must be ensured that the obligatory modifier is identified (in a case of ellipsis, it is necessary to take into account the preceding sentence, although often the Czech deletion goes in parallel with that in the corresponding Russian sentence).

4.1 One of the relevant differences between Czech and Russian syntax concerns sentences with the Czech 1st person plural corresponding to the Russian reflexive forms, e.g. Czech "Algoritmus rozmísťování bloků popisujeme v části 6" vs. Russian "Algoritm razmeščenija blokov opisyvaetsja v razdele 6" ("The algorithm of dislocation of blocks is described in Sect. 6"). Often a modal expression is present: "Názvy programů můžeme najít v knihovně" vs. Russian "Nazvanija programm možno najti v biblioteke" ("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:

$$\mathrm{Noun}_{acc}\ \mathrm{Verb}_{1stPl} \rightarrow \mathrm{Noun}_{nom}\ \mathrm{Verb}_{refl}$$

$$(\mathrm{Noun}_{acc})\ \mathrm{Verb}_{modal,1stPl}\ \mathrm{Verb}_{inf} \rightarrow (\mathrm{Noun}_{nom})\ \mathrm{Modal}\ \mathrm{Verb}_{inf}$$

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

4.2 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 lètě proběhlo jednání o nové variantě OS.
Russian:
Letom prošlo soveščanije o novom variante OS.
(In summer, the negotiations on the new variant of OS took place.)
b. Czech:
V lètě proběhlo jednání o prázdninách.
Russian:
Letom soveščanie prošlo vo vremja kanikul.
(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 vremja" with genitive.

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

c. Czech:
Práce na programu pokračují i v tomto roce.
Russian:
Raboty nad programmoj prodolžajutsja i v étom godu.
(The works on the program continue also this year.)
d. Czech:
Práce na fakultě pokračují i v tomto roce.
Russian:
Raboty na fakuľtete prodolžajutsja i v étom godu.
(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" ("whether"), "jestliže" ("if"), "když" ("when"), "dokud" ("till"), "dokud ne" ("until"), "pokud" ("as long as"), 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","pokud","jestliže" as Russian "esli"; Czech "dokud","dokud ne" as Russian "poka","poka ne", Czech "když" as Russian "kogda".

It follows that while it is necessary to work to a certain degreewith 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", "nachodit'sja", "imet'sja". 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 "nachodit'sja" is at place, otherwise the appropriate form of the copula is chosen; . Of course, another point concerns the translation of "být" within idioms ("byt' v porjadke", but "imet'sja v rasporjaženii").

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:
Étogo daže sistema točno ne znaet.
(This even the system does not know exactly.)
b. Czech:
Tabulka není uložena na pevném místě v paměti.
Russian:
Tablica pomeščaetsja ne na postojannom meste v pamjati.
(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 clitics in Czech, but usually follow the verb in Russian:

a. Czech:
... vypadal by tak, že by tabulka obsahovala údaje ...
Russian:
... vygljadel by tak, čto tablica soderžala by dannye ...
(... he would look as if the matrix contained(cond.) data ...)
b. Czech:
Budeme se v operačních systémech snažit ...
Russian:
V operacionnych sistemach budem starat'sja ...
(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 lexicon 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, Academia, Prague, 1980