Meta
Journal des traducteurs
Translators' Journal

Meta

Machine Translation Research in Czechoslovakia

Jan Hajič, Eva Hajičová and Alexandr Rosen

Volume 37, Number 4, décembre 1992

Études et recherches en traductique / Studies and Researches in Machine Translation

URI: https://id.erudit.org/iderudit/002996ar
DOI: https://doi.org/10.7202/002996ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Hajič, J., Hajičová, E. & Rosen, A. (1992). Machine Translation Research in Czechoslovakia. *Meta*, *37*(4), 802–816. https://doi.org/10.7202/002996ar

Article abstract

Machine translation research activities in Czechoslovakia starting in early the 60's are outlined, together with the basics of the theoretical background on which the parsing and representation levels have been based. Two more recent systems are described in more detail: APAC, working from English to Czech on INSPEC technical abstracts, and RUSLAN, which, translating from CZECH to Russian, was heavily taking advantage of the closeness between these languages. We conclude with a short description of the current project, which emphasizes the use of text corpora in combination with (more) traditional approaches. Many of the ideas we want to elaborate in the near future are present in the current project, and a word on future plans is also added.

MACHINE TRANSLATION RESEARCH IN CZECHOSLOVAKIA

JAN HAJIČ*, EVA HAJIČOVÁ, ALEXANDR ROSEN†
Charles University, Prague, Czechoslovakia

Résumé

Nous esquissons un panorama des recherches en traduction automatique en Tchécoslovaquie depuis le début des années 60, ainsi que des bases théoriques sur lesquelles sont fondées les analyses et les représentations. Sont ensuite décrits plus en détails deux systèmes plus récents : APAC, qui traduit de l'anglais vers le tchèque des résumés techniques de l'INSPEC, et RUSLAN, qui, traduisant du tchèque vers le russe, tire le meilleur profit de la parenté entre ces deux langues. Pour conclure, nous donnons une brève description du projet actuellement en cours, qui met l'accent sur l'utilisation de corpus en conjonction avec des approches (plus) traditionnelles. Un certain nombre des idées que nous voulons explorer dans un proche avenir sont présentes dans le projet actuel, et nous ajoutons quelques mots sur les projets à venir.

Abstract

Machine translation research activities in Czechoslovakia starting in early the 60's are outlined, together with the basics of the theoretical background on which the parsing and representation levels have been based. Two more recent systems are described in more detail: APAC, working from English to Czech on INSPEC technical abstracts, and RUSLAN, which, translating from CZECH to Russian, was heavily taking advantage of the closeness between these languages. We conclude with a short description of the current project, which emphasizes the use of text corpora in combination with (more) traditional approaches. Many of the ideas we want to elaborate in the near future are present in the current project, and a word on future plans is also added.

1. BACKGROUND

Machine translation research and development in Czechoslovakia is closely connected with the beginnings of the study of language from a computational perspective. Such activities date from the late 1950's, and resulted in the first experiments with English-to-Czech machine translation at Charles University in Prague in the early 1960's. Although the team, ideas and environment have changed considerably during the past 30 years, Charles University with the group around Petr Sgall still remains the centre of Czechoslovak research in this field.

The intertwinement of theoretical and practical aspects led to a rather realistic attitude towards machine translation (MT), which made it possible to avoid both the visions of a fully automatic high quality MT of the late 1950's and the consequent disillusionment in the 1960's. The first considerations in the domain of MT were guided by the effort to make use of the stimulating achievements of the Praguian linguistic tradition, especially with the intent to study the possibility of building up an intermediate language (Sgall 1963). In the long run, these considerations resulted in the proposal for the representation of (linguistic or literal) meaning, called *tectogrammatical representation* or

* Currently working as visiting scientist at IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA.

underlying structure. This proposal, together with the principles of a generative semantic base, was the basis of an alternative generative description of language, the so-called *functional generative description* (*cf.* its first formulation in Sgall 1964; for its more recent shape see Sgall *et al.* 1986). As we shall see below, the *tectogrammatical representation* was the framework for specifying the output structure of the English (and later Czech) analysis for MT. However, the main computational outcome of the theoretical investigations directed towards a generative description has been a broadly conceived and detailed system of random generation of Czech sentences.

2. APAČ — AN ENGLISH-TO-CZECH MACHINE TRANSLATION SYSTEM: PRINCIPLES VERSUS PRACTICAL SOLUTIONS

2.1. FIRST STEPS

After the first modest experiments with the English-to-Czech MT in the early 1960's, a new project was started in 1976. Work on this project, named APAČ (*Automatický překlad z angličtiny do češtiny* — "automatic translation from English to Czech"), was pursued until 1988, and later inspired a system for MT from Czech to Russian.

The APAČ project was actually a series of experiments. The first experiment, accomplished in 1978, was carried out in close cooperation with the TAUM group at the University of Montreal, which determined a number of its characteristics, most notably the transfer approach. Their major contributions, however, were Colmerauer's *Systèmes Q* formalism (Colmerauer 1970) and a version of an English analysis grammar designed by R. Kittredge.

The Q-systems, as a formalism for the definition and execution of a sequence of subgrammars, consists of unordered sets of rewriting rules which operate on a string of labelled trees. Analysis trees are built from the input string of individual words by the application of grammar and dictionary rules. Any case of homonymy (or synonymy in the synthesis) is expressed as parallel strings and results from several rules being applied to the same string. However, a Q-system outputs only the string(s) spanning the largest portion(s) of the input, as the most probable correct result. The parser is basically context-free and works bottom-up, parallel and all-paths.[1] The English analysis module was combined with an extensive and sophisticated synthesis of Czech, originally designed for random generation, into a test program which successfully translated several sentences taken from a newspaper text on economy.

In the next stage the analysis part was substantially modified by replacing the original constituency structure with the dependency approach. Another replacement was motivated by efficiency considerations: the Czech synthesis module was implemented in Q-systems as a special-purpose grammar.

The analysis module was able to parse much more than a few sentences, and it actually analysed a number of English sentences of various patterns. A simple morphemic analysis covered the regular inflected forms, syntactic rules identified the basic modifications of nouns and complex verbal forms, and a set of rules handled most of the elementary syntactic structures using valency frames from the lexicon. However, the dictionary coverage was quite limited, only one type of dependent clause (concessive clause) was treated, and only a modest attempt to solve idiomatic structures was made.

2.2. APAČ-2: THE STIMULATING EFFECTS OF TESTING THE SYSTEM ON REAL TEXTS

The second experiment (described in full detail in Kirschner 1982) differs from its predecessor in several respects, such as the introduction of elementary semantic features. For this experiment (called APAČ-2), the goal was the practical application of a batch-oriented MT system in information acquisition. The source texts, abstracts on microelec-

tronics, came from the *INSPEC* tape service. Even though the overall strategy resembles that used in the first experiment, several changes and improvements were introduced.

2.3. MORPHOLOGY

The program covered the entire system of English inflectional morphology, including almost all irregular, anomalous or rare forms. (The core of the procedure remains, however, true to that used by the TAUM group in 1973, with the kind approval of the authors). A few of the most productive derivational affixes are also included.

2.4. TRANSDUCING DICTIONARY

To reduce the constant need for new dictionary entries, especially from the ever-growing terminological vocabulary, the so-called *transducing dictionary* has been added. This module consists of rules which attempt to assign basic grammatical and semantic information to, and form a possible Czech equivalent for, some classes of terms of Latin and Greek origin, if such terms have not been found in the system's regular dictionary, *i.e.* if their grammatical and semantic properties are not idiosyncratic. The unknown word's final segment is identified and the whole word is transliterated according to the standards of Czech loanword orthography. Thus *application* is correctly translated by *aplikace*, *philosophy* by *filosofie*, *amplifier* by *amplifikátor* (although here the native *zesilovač* is more appropriate), *operational* by *operační*, even *rectify* by *rektifikovat*, or *privatize (privatise)* by *privatizovat*, etc. Words successfully identified in this fashion may be subject to the same set of morphological processes as the words found in the regular dictionary. Thus *privatize, privatizing, privatized, privatization, unprivatized, reprivatization* and also some other unlikely forms are all correctly translated, thanks to the closeness between English and Czech in this corner of vocabulary, of course.

2.5. COMPLEX NOUN PHRASES

Particular attention was paid to the syntactic analysis of nominal complexes in general and compounds in particular, especially with regard to the problems of correct structure assignment (in the English analysis) and part-of-speech conversion (in the transfer). A repertoire of semantic features gradually built on a highly schematic model of the universe of discourse helped to formulate rules that cover the regular, or at least the most frequent phenomena in the domain. The following example illustrates the mixture of monolingual (English) and translational syntactic ambiguity potentially present in the phrase *an integrated circuit system*, which has to be translated in one of the following ways:

(a) *systém integrovaného obvodu*
"a system of an integrated circuit"
(b) *systém integrovaných obvodů*
"a system of integrated circuits"
(c) *integrovaný systém obvodu*
"an integrated system of a circuit"
(d) *integrovaný systém obvodů*
"an integrated system of circuits"
(e) *integrovaný obvodový systém*
"an integrated circuit system"

None of the possible translations preserves the structural ambiguity of the English phrase, and all except (e) introduce the translational ambiguity of number for *circuit*. To select (b) as the correct equivalent, first the choice is limited to (a) or (b) by identifying *integrated circuit* as a stable collocation, and then the rules infer from the semantic features of *system* and *integrated circuit* that the genitive attribute should be plural.

2.6. COORDINATION

Since the source texts were rich in coordination structures, several sets of rules were included, aimed at analyzing different types of conjunction at both the phrasal and sentential levels.

2.7. INTEGRATING TRANSFER WITH ANALYSIS

Radical steps were taken to direct the analysis more towards the target language and, as a major deviation from a strictly transfer-oriented approach, to eliminate a separate transfer phase. Wherever possible, Czech lexical equivalents are already supplied at the initial stages of the English analysis; sets of indices required for the synthesis of Czech are already introduced during the analysis of English from the stage of dictionary lookup on; lexical information for the analysis, such as English semantic features or valency frames, is always deleted as soon as it has fulfilled its task.

2.8. NO SURRENDER OF THEORETICAL SOUNDNESS

Although the development of the system tended towards application-specific solutions and immediate results, the authors tried to retain a preference for a more general approach. *E.g.* the system restores those elements deleted in the surface structure of a sentence which are necessary for its semantic interpretation (and thus also for the proper choice of its equivalent construction in Czech).

2.9. PRESERVATION OF AMBIGUITIES

More general solutions were also preferred to meet a problem specific to the task of English-to-Czech translation. In English, due to its poor morphology offering only limited means for expressing referential relationships via grammatical concord, extralinguistic knowledge plays a more important role than in Czech, where elements bound together by referential relationship must agree in case, gender, number, and, with verbs, in person. A system without such knowledge will face difficulties if it is to decide to which nominal complex the verbal attribute *using* belongs in a sentence such as *These methods employ a Monte Carlo analysis in the parameter space using a simplicial approximation to the region of acceptability*. A layman can exclude *space* as an agentive by the same means as the system, which can also be endowed with the knowledge that, under normal circumstances, *space* cannot *use* anything. However, for the other two candidates *methods* and *analysis* the decision will be difficult without at least some idea of what *Monte Carlo analysis* and *simplicial approximation* are. A practical system cannot produce routinely multiple results. Lacking tools to implement computationally more sophisticated methods for selecting more probable outputs from a number of suggested alternatives, the authors decided on a policy to prevent the occurrence of ambiguities in the first place. If the target language equivalent can be made as ambiguous as the original utterance, then the correct interpretation can be left to the reader and the ambiguity can be suppressed even in the source language analysis. This solution is preferred even if the ambiguity is real, as in the above example, and not just "translational". It is preferred even if the equivalent does not fit stylistically. So *using*, as well as a number of other verbs occurring in the above pattern, is not translated by a verb which would have to show agreement (in Czech even in the transgressive form, though only in gender). *Using* is converted into a prepositional phrase *with* + deverbal noun (*s použitím*) and the nominal complement is marked by the genitive case. The English paraphrase of the result could then be *with the use of a simplicial approximation*.

2.10. FAIL-SOFTNESS

To make the system work in other than laboratory conditions, it was realized that means are necessary for coping with phenomena that go beyond the system's current state of development. For the simplest case when a word is not identified by any of the dictionary operations, a "universal" noun is supplied with a "universal" set of semantic features, retaining its original lexical value. A subgrammar is aimed at solving syntactic and other anomalies as well as the system's inadequacies, which would otherwise result in failures. *E.g.* if a noun is used as a metaphor, the intersection of the set of its semantic features and the set of features required in the respective slot of the valency frame remains empty, the noun fails to become integrated in the verbal complex and the sentence is not analysed as no mechanism for releasing the conditions is available. In Czech translation, the verb would lack agreement marking and its complements would lack case (or prepositions) required by the verb's subcategorization. However, when defaults are applied to give the most frequent values for the missing morphological categories, the result improves considerably.

2.11. THE RESULTING STATE

The development of the system continued throughout the 1980's, leading to gradual expansion of the lexicon (which, however, did not exceed 5,000 entries), a preference for richer and more compact lexical entries with information structure favouring systematic and general treatments rather than ad hoc solutions, a larger coverage of grammar phenomena, a grammar more finely tuned to the input texts, etc. The system was also tested on abstracts of technical literature on pumps and other hydraulic devices, a domain slightly easier when compared to electronics. However, the original plan of a practical application was finally abandoned and the project phased out in the late 1980's.

In its final version, the system embodied the following ideas from the theory of *functional generative description*:

(a) representation of grammatical relations in terms of dependency structures;
(b) stratification of the processes of analysis and synthesis and their results into levels corresponding roughly to morphemics, surface syntax and tectogrammatics (deep structure);
(c) articulation of the distinction between function and content words;
(d) the concept of the level of meaning (tectogrammatics) as the highest level within a language system.

The following example sentence should give a picture of the system:

(1) The carpet plots so obtained permit the best operating conditions for each type of aerofoil to be immediately identified and thus the most suitable section can easily be selected for a given specification.

The Czech translation provided by the system is as follows:

(2) Tak získané kobercové diagramy dovolují, aby byly
so obtained carpet-like plots permit that were
bezprostředně identifikovány nejlepší provozní
immediately identified best operating
podmínky pro každý typ profilu, a nejvhodnější
conditions for every type of profile and most suitable
část může být tak snadno vybrána pro danou
part can be thus easily selected for given
specifikaci.
specification.

Improvements can certainly be made, but the sentence preserves the meaning and is understandable, which is sufficient for the purpose of information acquisition. However, before this result was obtained, the system translated *the best operating conditions* as *nejlépe fungující podmínky* (*i.e.* "conditions operating best"). While it might be risky to prohibit *condition* and similar nouns as the actor of *operate*, the solution was trivial: *operating condition* was treated as a stable collocation, with the additional benefit of easy selection of the best-fitting Czech equivalent.

The analysis tree of the sentence is expressed in the Q-systems formalism in the following way (the printout is edited and English lexical values are retained to enhance readability):

```
S(V(COOR(and,/),
   V(permit(/),
     N(plot(L,$AG,*DEF,*C,*A,*VB,/,*PL),
       AD(obtain(L,$ATR,*VAD1,/),
           ADV(so(L,$ADV,PSC,/))),
       AD(carpet(L,$ATR,*C,/,*SG))),
     V(identify(R,$ADV(*PURP),*INF,*MNR,*RFX,*PSV,/),
       ADV(immediate(L,$ADV,*MNR,/),
       N(NIL(L,$AG)),
       N(operating-condition(R,$PAT,*DEF,*A,*COLL,/,*PL),
         AD(good(L,$ATR,*SUP,/))),
       N(type(R,$ADV(*PURP(for)),*C,*A,/,*SG),
         AD(every(L,$ATR,/)),
         N(aerofoil(R,$ATR(of),*C,*A,/,*SG)))),&)),
   CONJ(and),
   V(select(*MNR,*PSV,*CAN,/),
     ADV(thus(L,$ADV,*EMP,/)),
     ADV(easy(L,$ADV,*MNR,/)),
     N(NIL(L,$AG)),
     N(section(R,$PAT,*DEF,*C,*A,/,*SG),
       AD(suitable(L,$ATR,*SUP,/))),
     N(specification(R,$ADV(*BENF(for)),*IDF,*A,/,*SG),
       AD(give(L,$ATR,*VAD1,*VPS,*AUTH,/))))))
```

2.12. COMMENTS

2.12.1. LINGUISTIC THEORY: DEPENDENCY-BASED GRAMMAR

Throughout the APAC project, dependency-based structures are used, implemented as trees in the Q-systems formalism. The trees, as the only type of structure allowed in the Q-systems, serve the double purpose of representing the dependency relations and structuring all data relative to a content word. A subtree with a word class symbol at its root is used to represent the analysis of a content word of the terminal string. Besides the word class of the word, the subtree gives the following information:

- the word's lexical value
- its position in the (deep) word order relative to the governor (L = left, R = right)
- the symbol for the (tecto) grammatical function ($AG = actor, $PAT = patient, $ATR = attribute, $ADV = adverbial)
- other word-class-specific symbols, *e.g.* for definiteness and indefiniteness (*DEF/*IDF), for elementary semantic classification (*C = concrete, *A = abstract, for word-class homonymy (*VB = noun-verb), for grammatical number (*SG/*PL), etc.

Coordinated elements are represented as daughters of a non-terminal node marked COOR.

2.12.2. LEXICON: VALENCY FRAMES AND SEMANTIC FEATURES

The dependency-based analysis allows for easy employment of frames which restrict not only the choice of a verb's complements as proper valency frames, but are used also for the interpretation of typical or somehow idiosyncratic optional modifications of verbs, nouns, adjectives and adverbs, especially by prepositional phrases. Moreover, every adjective has a single-slot frame requiring certain semantic features of its head noun. Another component of the frames is the information concerning the constitution of compound verbs, *i.e.* phrasal and prepositional variants of verbs. Frames also contain information on semantic features, *i.e.* selectional restrictions, which usually are, after some bitter experience, quite permissive. Four basic groups of semantic features are used:

(a) features that help recognize a metatextual framework in the abstracts (*e.g. the paper describes, discusses, treats*, etc.);
(b) features distinguishing terminological expressions and reflecting the position of individual terms in the system (*e.g.* most general categories, expressions that occur both in a general and in a field-specific meaning);
(c) features indicating general conceptual properties (abstract, concrete, human, action, property, etc.);
(d) features connected with the role or function of the denoted objects in the particular field (material, instrument, location, etc.).

These features represent the properties of their bearers as well as restrictions imposed by heads upon their slot fillers.

The following sample of lexical entries should illustrate the use of frames and semantic features, as well as the true bilingual quality of the whole system (see Linguistic Techniques, below). The entries are not edited, except that shorthand notation (*i.e.* macros) is expanded:

CARPET == N(KOBERC(MO5),*C).
N = noun
KOBERC = stem of the Czech equivalent
MO5 = Czech inflection paradigm
*C = concrete

SUITABLE == AD(VHODNE2(3),1(*A,*C,/),
*MNR,*M,TO(K3),FOR(PR04)).
AD = adjective
1(...) = slot for a modified noun, should be concrete (*C) or abstract (*A)
*MNR = derived adverb is an adverb of manner
*M = comparison by *more* and *most*
TO(K3) = dependent prepositional phrase with the preposition *to* is translated by the preposition *k* with the noun in the third case, *i.e.* dative

SELECT == V(VYBI2R(50I11,VYBER),1(*A,*C,*H,/),
2(*OB,*A,*C,/),*FIN,FROM(Z2),6).
V = verb
50I11, VYBER means that the stem VYBI2R inflects according to the paradigm 50 and is imperfective, the perfective counterpart is VYBER and inflects according to the paradigm 11
1(...) = the actor can be of any semantic category (abstract, concrete or human) while the patient —

2(...) is *OBligatory and cannot be human
*FIN = an infinitival clause following the verb is translated as a finite clause of purpose; a *from* prepositional phrase becomes *z* + genitive
6 = a participle derived from the verb should preferably be translated by the imperfective variant.

GIVE == V(GIVE,1(*A,*C,*H,/),2(*OB,*A,*C,/),3(*C,*H,/),
YPTC(VZESTUP,TO),YPTC(ZROD,TO),
*AUTH,*VPS,6,*IO,*REL1).

The translation of this verb is given in the synthesis (due to the possiblity of compounding)
3(...) = restricts an optional addressee
*IO = indirect object, *i.e.*, addressee at the underlying level; could be also effect or origin
YPTC(...) concerns the compound verbs *give rise to* and *give origin to*, which are treated as wholes to be translated later in the synthesis (the nouns already have their Czech lexical values at the stage where the compound verbs are identified)
*AUTH = the verb is often used in the metatextual context (*the author gives an example*)
*VPS = in most relevant contexts, the Czech equivalent of this verb prefers one of the two possibilities of passive voice formation, *i.e.* the complex passive over the reflexive passive, *e.g. an example is given — je dán příklad* (not *dává se příklad*)
*REL1 = in most relevant contexts, the *ing* form of the verb following a noun is the noun's attribute and can be translated as a relative clause; this feature blocks the analysis of the *ing* form of the verb as a deverbal noun, with the preceding noun as its modifier.

2.12.3. LINGUISTIC TECHNIQUES: QUASI-DIRECT

The approach can be characterized as direct in the sense that the system was designed from the start to translate from English into Czech (and not *e.g.* into Hungarian). This, however, has nothing to do with simple one-by-one substitution of elements or structures.

The analysis bears characteristics independent of the target language, such as full morphological analysis and, in most cases, identification of the underlying relations. It might be possible, although not quite straightforward, to adapt the analysis for other, preferably related languages (*e.g.* Russian).

The most conspicuous manifestation of the "directness" of APAČ-2 (as opposed to the transfer approach of APAČ-1) is the assignment of Czech equivalents of English words and structures as soon as possible, *i.e.* already during the first lexical look-up and in the English analysis. A typical lexical entry in the analysis lexicon gives all the data also required also for transfer-like operations and synthesis (see above). However, the possibility of performing lexical or structural selection later in the synthesis is always open, if the correct equivalent cannot be determined earlier than that. This concerns *e.g.* compound verbs and nominal collocations, some cases of word order transformations, and, quite prominently, the decision on verbal aspect, since an English verb typically corresponds to at least two different verbs (*i.e.* surface lexical units) which explicitly mark aspect for every form. Thus the Czech infinitive of *select* can be either *vybírat* or *vybrat*.

2.12.4. FAIL-SOFT (EMERGENCY) MEASURES

The emergency measures applied by APAČ-2 serve the following purposes:

(a) They prevent the program from being stone-walled or otherwise abnormally terminated.
(b) In the analysis, they interpret unrecognized units (mostly word-level units, but occasionally also partial parses) and integrate them into more complex structures.
(c) Whenever possible, they form Czech equivalents for the unidentified units. For lexical units, this is done either by adapting words recognized as internationalisms (see 2.1. above on "transducing dictionary"), or by "czechizing" English words,

i.e. assigning them features and forms proper to their presumptive Czech counterparts, such as part of speech, gender and a Czech suffix.

The system of emergency measures thus (1) treats elements not found in the standard lexicon and (2) remedies failures to arrive at an accomplished parse. In the latter case, the synthesis can process even partial results and attempt to compile some output, with a warning for the reader that the translation was produced by non-standard means and that the distance between the meaning of the original and its translation can be substantial.

Finally, if everything goes wrong and only a string of words and dubious partial parses is produced, a set of "sweeping rules" is invoked to polish the output to be at least readable. These rules are a constitutional part of the system's "preferential tactics".

2.12.5 PARSING STRATEGY: PREFERENTIAL AND AMBIGUITY-PRESERVING

As a device for implementing context-free grammars in an orthodox fashion, Q-systems lack explicit means to control the parsing process. Therefore, any attempt to favour some parses or translations from a number of grammatically correct, albeit not equally acceptable, choices is a rather painful enterprise. Instead of a system of weights or some other (maybe more appropriate) method, the necessary preferences are implemented by repeating at later stages of processing a set of the most important rules in a more "liberal" version with looser conditions. Thus, more likely solutions (giving relatively "safe" partial parses) are produced earlier than those corresponding to rarer phenomena (and larger wholes), which are treated later by less restricted rules. This, of course, works only if local context does not deceive the program into discarding a correct but unlikely interpretation by favouring a wrong one at some earlier stage.

This approach is closely connected to the problem of ambiguity resolution. Recognized instances of ambiguous structures regarded as decidable are treated before they can cause a process leading to multiple parses of the whole sentence. Cases which cannot be decided at an early stage of the analysis are kept implicit until they can be resolved, *i.e.* until the synthesis. Only cases undecidable in this fashion produce all relevant combinations. Such a strategy helps to avoid parallel parses that give rise to identical structures in the target language, and thus can be said to preserve the ambiguity of the corresponding source language structures (for a more detailed discussion, with rich material illustrating different types of ambiguities, see Kirschner, 1987).

2.13. RESULTS AND CONCLUSION

After substantial increase of its lexical coverage, the system could serve the purposes of information scanning in a given text domain; the results would require post-editing, and, should more than information-scanning be pursued, a profound expert revision. The quality of the output varies with the sentence length and complexity. Excessively long or very intricate constructions, especially those overloaded by parallel meanings, may even cause failures. "Blind" tests with unknown words usually end up with 20% translated adequately (with the unknown words treated by auxilliary devices), 20% requiring translation from scratch and the rest suitable for revision. The results for texts with no unknown words show less percentage of rejected input and better quality of the bulk worth post-editing.

The speed is determined by the input in a similar way[2].

The reasons why the original goal of practical application was not attained were partly due to the unrealistic expectations of the sponsors (and future users) and the resulting lack of progress in expanding dictionary coverage, but the inherent limitations of the system itself probably played the major role. The continuity trend in the system's development

towards solving all transfer problems within the analysis phase was beneficial as a short term measure, but it paid a high price in the long run. A single rule was often insufficient to treat English structure: the rule also had to decide which Czech equivalent would be used and then it had to be split accordingly into more versions. The English analysis gradually developed into a very complex program whose maintenance became a major problem. Since not everything could have been solved this way, the data were often "bilingual" and the rules had to take this into account. And, as mentioned above, the risk involved in implementing preferences as subsequent loosening of restrictions based on local context manifested itself in frequent misinterpretations blocking correct analysis. A more principled approach, perhaps with a distinct transfer and a "clean" analysis, plus a better mechanism for stating preferences, might have improved the system's chances.

3. RUSLAN — A CZECH-TO-RUSSIAN MACHINE TRANSLATION SYSTEM: NO SHORTCUT TO SUCCESS WITH RELATED LANGUAGES

3.1. ORIGINS AND SOURCES

In 1985, a project for machine translation of Czech software manuals into Russian was started as a second MT project of the group of mathematical linguistics at Charles University. Work on the project continued until 1990.

The goals were both practical (translation or re-translation of new or re-edited manuals for export purposes within the former COMECON countries, of an estimated amount of 500 to 1,000 pages a year) and theoretical (test of an approach to the analysis of Czech and development of a theoretical background for MT between closely related languages). The project was funded by the software producer, the Research Institute of Mathematical Machines (VÚMS), Prague, and carried out in cooperation with the university group at the Faculty of Mathematics and Physics.

One pleasant feature of the project was that the input texts were quite predictable in the sense that almost all of them were available already at the time the system was being developed. They described components and utilities of an operating system named DOS-4 developed at VÚMS as an advanced extension to the common DOS. Approximately 70 manuals with about 13,000 pages, 1,700,000 running words and 54,000 different word forms were available on magnetic tape for testing and pre-processing. They were maintained under an editing/formatting system PES (Programmed Editing System), which supports preparation, editing and binding-ready printing using national characters for Czech and Russian. The texts were kept in an internal format with editing and formatting commands, version identification, information on last-changed pages, etc. Whereas most of this could be used to improve the system's efficiency, some data had to be handled with care.

3.2. THE OVERALL STRUCTURE

RUSLAN is a unidirectional bilingual system. The translation scheme is transfer-like in the sense that no intermediate pivot language is used. However, many simplifications were made with the expectation that the close relationship between Czech and Russian does not require a full-fledged transfer scheme. The result of this is that the system resembles in some respects the so-called direct method.

Translation proceeds automatically in batch mode without human intervention during the process. The ambition was to obtain high-quality results which would require a minimum of post-editing, comparable to human translation. No manual pre-editing was expected.

The translation unit is a single sentence. Thus, the recognition of sentential boundaries is a part of preprocessing.

The following steps are performed during the process of translating a given (part of a) manual:

1. the text is extracted from the tape, to "visualize" all emdedded editing and formatting commands;
2. fully automatic preprocessing: — national and special characters conversion and coding — sentence boundaries recognition;
3. Czech morphological analysis;
4. Czech syntactico-semantic analysis with respect to the Russian sentence structure, for each input sentence;
5. conversion of the analysis result into a string of annotated Russian base forms; some transfer-like operations are performed at the same time;
6. morphological synthesis of Russian plus integration of preserved editing and formatting commands into the result;
7. the output is saved onto a tape under the PES system again.

The resulting text can then be easily printed and corrected using PES editing tools.

3.3. MORE DETAILS

3.3.1. PREPROCESSING

Words and punctuation symbols are distinguished from other information in the source text which does not require translation. Special characters (such as mathematical symbols and Greek letters) as well as PES commands are encoded and attached to the nearest "real" word.

To recognize sentence boundaries a special algorithm was developed, which takes into account both editing commands and punctuation with upper/lower case shifts. This was the most challenging part of this step.

3.3.2. MORPHOLOGICAL ANALYSIS

Morphological analysis plays quite a substantial part in this system, because of the sheer complexity of the Czech morphology, and because of the richness of the information provided for the syntactico-semantic analysis.

The morphological analysis is based on pattern unification. The dictionary look-up provides all possible stems; ambiguities are treated in parallel during the next step.

For words which were not found in the dictionary, a procedure similar to that employed in APAČ ("transducing dictionary") was implemented: unidentified words whose final segments suggest a Latin or Greek origin are assigned a relevant set of features, the final segment is adapted and the rest of the word is transliterated according to the standards of Russian. The idea to transduce also words which are related in Czech and Russian by their common Slavonic origin was soon abandoned because of their frequent semantic and morphological incompatibility.

3.3.3. SYNTACTICO-SEMANTIC ANALYSIS

This is the most important step. For the theoretical background of the approach, see Sgall *et al.* (1986); for a detailed description of the actual analysis program with many examples, see Oliva (1989). The core of the analysis was based on the results of an earlier project, TIBAQ (Text- and Inference-Based Answering of Questions, see Hajičová and Sgall, 1980, and Panevová and Oliva, 1982), namely on its independent Czech analysis for the purpose of automatic understanding of written texts.

The cornerstones are the same as those of APAČ: dependency grammar and data-driven parsing with heavy reliance on valency frames. To control the combinatorial

expansion, semantic features are used as additional constraints. This system, however, relies much more on the lexicon with the effect that the grammar rules can be more general and the grammar more compact and easier to maintain. The Q-systems formalism with its bottom-up and all-paths parsing strategy was found very useful for the analysis of Czech as a free word order language.

As in APAČ, the analysis produces structures which are not representations of the Czech input, but rather of its Russian translation. Therefore, no separate transfer step was needed. Most of its task is performed in the analysis, where a typical rule analyzes and transfers at the same time.

Furthermore, due to the close relationship between Czech and Russian, many ambiguities can be left unresolved, because any attempt to disambiguate would eventually lead to a number of identical surface strings in Russian.

For other cases of ambiguity, where the translations are different or where no reading was determined as being by far most frequent, multiple outputs could not be avoided. However, keeping the post-editor's task in mind, all disjunctions in the final output are made as local as possible. Thus, lexical ambiguities are not allowed to multiply the number of translations of a single sentence and the preferred reading is highlighted.

3.3.4. GENERATION

While the dependency tree is being decomposed, morphological information is sent from the governor to its dependents according to agreement and government specifications. The original word order is slightly altered, if necessary. An ordered list of word stems with morphological information and the restored editing/formatting attributes is the output of this step.

3.3.5. MORPHOLOGICAL SYNTHESIS AND DECODING

The module processes word stems with morphological information to obtain their inflected forms. A limited amount of derivational morphology is also included (*e.g.* deverbatives). The module is also responsible for orthographical changes when some prepositions and pronouns precede certain words.

After synthesis, each word with its attributes is decoded into the PES-acceptable format. This is an inverse operation to step 2.

3.4. IMPLEMENTATION

Steps 1 and 7 are handled by special software, which is a part of the operating system DOS-4. Steps 2 and 6 are written in standard Pascal (including the module of Russian morphological synthesis). Steps 3 to 5 are grammars expressed in the formalism of Q-systems[3].

The maximum memory requirement is 640 kB. Secondary storage volume depends on the size of the lexicon: an average entry occupies 300 bytes. In the last version, there were 10,000 lexical entries.

Elapsed time needed for translation depends on hardware and the time sharing coefficient. The fastest version of EC-1027 translated an average word in less than 3 seconds CPU, which would suffice to provide the desired output of 50 pages a day.

We also performed a small-scale (700 sentences) evaluation of the results only a few months before the project ended. We used the standard *grammaticality* and *intelligibility* scales, with the overall conclusion that about 20% of the translations were correct, about 20% had to be translated from scratch again, the quality of the rest was somewhere in between. From a practical point of view, however, the amount of postediting required (measured in time) to get these translations in order was already only slightly higher than that for revisions of translations produced previously by human translators (on the same manuals).

3.5. CONCLUSION

There were several reasons why the original plans to use the system for routine translation were abandoned.

Undoubtedly the crucial one was, rather surprisingly, the changes in the former COMECON countries. The artificially sustained demand for an obsolescent technoloy turned towards cheaper and better imports from the newly opened world. Furthermore, the Soviet market collapsed. There was no longer a need for translating software documentation for Russian customers, bacause there were no longer any. The regulation demanding Russian translation of any documentation related to software products for the COMECON market was no longer enforced.

But there were other reasons as well, which would have probably made their impact on further progress of the project. All of them had to do with a certain underestimation of the problems involved in machine translation, even between related languages and within a restricted text domain.

The primary source of problems with the texts were the frequent occurrences of phenomena which traditionally escape the attention of linguistics: the texts were full of non-sentences (lists, mathematical expressions), punctuation symbols (embedded brackets, dashes, etc.), and English or English-like quotations from programming languages and non-standard terminology. It was difficult to find exhaustive solutions to all such cases. Furthermore, it was found that the original estimation concerning the grammar structures present in the texts was too optimistic and that the coverage of the Czech analysis grammar had to be expanded.

However, the main obstacle to a sizable improvement of the translation quality had to do with the underlying assumption that transfer is a negligible part in a system for Czech-to-Russian translation of technical texts. Ad-hoc rules had to be used for rather common phenomena. The lesson was that even for related languages one should not avoid principled solutions.

4. MATRACE — MACHINE TRANSLATION BETWEEN CZECH AND ENGLISH: EXTRACTING A CONTRASTIVE KNOWLEDGE FROM BILINGUAL TEXTS FOR TRANSFER IN A NEW SYSTEM

4.1. THE RATIONALE

In 1989/1990, after the past two projects, in a situation where future research funding was a major concern, a simple system of machine-aided English-to-Czech and Czech-to-English translation built around an on-line dictionary, runnable on PC-compatibles, with possibilities open for further development, seemed like a good choice. The situation changed, however, when an offer came to make the machine translation research part of the IBM Academic Initiative project in Czechoslovakia. Considering the worldwide trend towards the use of machine-readable text corpora for both theoretical and practical purposes, the virtual nonexistence of electronic sources of linguistic data for Czech, and the problem of acquiring enough data for transfer by traditional methods, the decision was made to base a future machine translation system on knowledge extracted from correspondences between parallel bilingual linguistically tagged texts.

4.2. THE GOALS

The project started in late 1990 and its first stage is concerned with the extraction of lexical and structural correspondences from large parallel Czech and English texts, morphologically and syntactically analysed. The long-term goal of the whole entreprise is the development of a machine translation system between Czech and English. The first stage will provide data which should later be absorbed and used by a module of lexical

and structural transfer between the two languages. Individual words, groups of words and whole sentences from texts in both languages will be annotated by syntactic parsers with morphological and syntactic categories as well as structural information and, if necessary, hand-corrected. The results of this linguistic analysis will be presented within a simplified version of the dependency framework of the *functional generative description* and used later for the (semi-) automatic generation of structural and lexical correspondences between grammatical units of the two languages. These correspondences should then become the basis for a transfer module in a machine translation system, the long-term goal. The results should prove the feasibility of the approach to machine translation where linguistic and statistical methods are applied together, and the practicality of the linguistic theory in the contrastive description of large volumes of text.

Besides the primary goal of a machine translation system, the collection of (bilingual) texts annotated by morphological and syntactic information will be available for text-oriented linguistic research. Tools such as implemented grammars will hopefully be usable in other applications and research environments.

The following results should be produced by the end of 1992:

(a) Czech and English monolingual dictionaries of reasonable coverage (English: 50,000 lexemes, Czech: 80,000 lexemes), usable for morphological and syntactic analysis of the texts;
(b) morphological analysis modules for Czech and English (morphological taggers);
(c) shallow syntactic analysis modules for Czech and English (syntactic taggers);
(d) tools for displaying and editing the analysis results;
(e) a tool for the alignment of parallel text units within the bilingual corpus;
(f) rudimentary tools for comparative studies of the tagged corpora.

4.3. THE CURRENT STATE

Work on the project is carried out jointly by two university institutes: the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics (head: Eva Hajičová) and the Institute of Theoretical and Computational Linguistics at the Faculty of Philosophy (head: Petr Sgall).

The primary source of lexical data for the first approximation of English analysis is the *Expanded Computer Usable Version* of the *Oxford Advanced Learner's Dictionary of Current English* (see Mitton 1986; 37,500 base form entries) in an adapted form. The Czech lexicon is based on a dictionary of 80,000 base forms with morphological information (for both inflection and derivation), originally compiled for a spelling checker, which is currently being augmented with syntactic valency frames for verbs.

Modules for the morphological analyses of Czech and English are ready and preparations are made for implementing the syntactic analyses. The Czech side will, at least provisionally, use a modified version of the Czech analysis grammar from the Czech-to-Russian MT project. The English syntactic analysis will be implemented in a constraint-based formalism and possibly supplemented by statistical disambiguation. A limited amount of machine readable Czech and English parallel texts is now also available.

5. ACKNOWLEDGMENT

This paper draws substantially upon the work of other people involved in the projects, especially upon the work of Zdeněk Kirschner, the initiator and the main author of the APAČ series.

Notes

1. The Q-systems compiler and interpreter are used with the kind permission of Benoit Thouin, the author.
2. However, average data can be presented. As to the hardware, APAC-2 has been implemented on computers of the size and type of IBM 360 or 370 (EC 1040, EC 1055, EC 1027, ICL 4-72, SIEMENS 7755). An average sentence of 15-20 words translates in about 1.8 minutes, the treatment of one word taking approximately 6 seconds. More recent implementations on IBM PC AT compatible machines take a comparable time.
3. The system was originally intended to run under operating system DOS-4, on EC-1027 or IBM/370 systems. However, during the development the system was made portable to IBM PC compatibles, with the exception of the steps 1, 2 and 7.

BIBLIOGRAPHY

COLMERAUER, A. (1970): *Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur*, Montréal, Université de Montréal.

HAJIČOVÁ, E. and P. SGALL (1980): "Linguistic Meaning and Knowledge Representation in Automatic Understanding of Natural Language", *The Prague Bulletin of Mathematical Linguistics*, 36:5-51.

KIRSCHNER, Z. (1982): "A Dependency-Based Analysis of English for the Purpose of Machine Translation", *Explizite Beschreibung der Sprache und Automatische Textbearbeitung IX*, Praha, Matematicko-fyzikální fakulta, University Karlovy.

KIRSCHNER, Z. (1984): "On a Dependency-Based Analysis of English for Automatic Translation", *Contributions to Functional Syntax, Semantics and Language Comprehension*, P. Sgall (ed.), Praha, Academia, pp. 335-358.

KIRSCHNER, Z. (1987): "APAC3-2: An English-to Czech Machine Translation System", *Explizite Beschreibung der Sprache und Automatische Textbearbeitung XIII*, Praha, Matematicko-fyzikální fakulta University Karlovy.

KIRSCHNER, Z. and A. ROSEN (1989): "APAC — An Experiment in Machine Translation", *Machine Translation*, 4:177-193.

MITTON, R. (1986): "A Partial Dictionary of English in Computer-usable Form", *Literary and Linguistic Computing*, 1:214-215.

OLIVA, K. (1989): "A Parser for Czech Implemented in Systems Q", *Explizite Beschreibung der Sprache und Automatische Textbearbeitung* XVI, Praha, Matematicko-fyzikální fakulta, University Karlovy.

PANEVOVÁ, J. and K. OLIVA (1982): "On the Use of Q-Language for Syntactic Analysis of Czech", *Explizite Beschreibung der Sprache und Automatische Textbearbeitung* VIII, Praha, Matematicko-fyzikální fakulta, University Karlovy.

SGALL, P. (1963): "An Intermediate Language in Machine Translation", *Proceedings of 26th Annual Meeting of the American Documentation Institute, Chicago, 41f*, Also *Computational Linguistics*, Budapest 1964, 2:35-62.

SGALL, P. (1964): "Generative Beschreibung und die Ebenen des Sprachsystems", Presented at the Second International Symposium in Magdeburg, Germany, printed in Zeichen und System der Sprache III, Berlin 1966, 225-239.

SGALL, P., L. NEBESKÝ, A. GORALČÍKOVÁ, E. HAJIČOVÁ (1969): *A Functional Approach to Syntax in Generative Description of Language*, New York.

SGALL, P., E. HAJIČOVÁ, J. PANEVOVÁ (1986): *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, co-edition Praha, Academia, Dordrecht, Reidel.