T. D. Crawford

A REVIEW
OF THE CARDIFF MACHINE TRANSLATION PROJECT

The aim of the Cardiff Machine Translation Project was the development of a fully automatic system for the mechanical translation into English of modern non-literary Italian texts. The project was conceived by Dr. Sparaco Gamberini of University College, Cardiff, and was to form the subject of a doctoral thesis on my part. Dr. Gamberini acted as my supervisor for the thesis, and contributed the formalized Italian grammar which was incorporated in the system.

The Italian text to be translated must first be prepared on punchcards so that it may be input as data on which the MT program can operate. No pre-edition of any kind is permitted, except that the user may, if he so wishes, insert at the beginning and end of each paragraph a symbol which will cause the machine to output the English translation with the original paragraph divisions preserved. For research purposes we established a convention that a space should be left before as well as after punctuation marks. This is not in any way essential, as the system could easily be redesigned to operate on the basis of the normal typing convention, which attaches punctuation marks to the end of the preceding word. But the mechanical separation of word and punctuation does necessarily consume a certain amount of computer time, and the convention we at present use is therefore more efficient. The accents used in Italian are not available on the keyboards of most punching machines, and we replace these by an apostrophe immediately following the accented letter.

The program itself is written in FORTRAN 4. This has the advantage of being a widely used language, which means that with very minor modifications our program can be transferred from one machine to another. The disadvantage is that FORTRAN 4 is not well suited to linguistic computation, but we have not so far found it such an obstacle as to induce us to use another programming language.

The machine operates on the input text one sentence at a time.
All the words occurring in the text are listed, and a search is then made in a mechanical dictionary to see if these words are included. Originally we had only magnetic tapes available for storing the dictionary, and since random access is impossible in the case of a tape-file, retrieving information from the dictionary was a lugubriously slow process. We therefore decided that for the purposes of the initial research we would confine our dictionary entries to those words actually appearing in the texts under study, and leave the compilation of a complete dictionary until we had more satisfactory hardware. In fact we now have the possibility of temporary storage on discs, which can be accessed at random, so that we are able to store a full dictionary on tapes, and copy it on to a disc before operating the MT program. Words not appearing in the dictionary are treated as proper names and not translated. For those words which it does find in the dictionary, the computer retrieves and stores temporarily in its core certain grammatical information and a provisional translation into English.

When the dictionary search is completed, the next step is the resolution of any homographs which may be present in the sentence. Words which must be translated in different ways depending upon the context are marked in the dictionary, and thus noted if any are present among the vocabulary searched for at the previous stage. We have devised three types of test which may be applied in order to decide which translation should be assigned to a homograph: I shall refer to these tests as 1) grammatical; 2) micro-contextual; 3) macro-contextual.

The grammatical test I shall discuss in connection with the Italian grammar, as it consists of trying each alternative version of the homograph to see which fits most readily into the syntactic context. It is a recent innovation in the program, and was not used to obtain any of the sample translations which have previously been made public.

The micro-semantic test exploits the fact that in the case of many homographs, one or more of the words occurring in the immediate context may give a clue as to the correct translation. For instance, the preposition da in Italian has by and from as its most frequent English translations. One may be fairly sure that if da is closely preceded by, for example, a form of desumere, detrarre, or derizzare, the correct translation will be from. Therefore we initially marked words of this type in the dictionary, and specified that if da occurred in the sentence to be translated, and one of these marked words closely preceded it, da should be translated as from; otherwise it should be translated as by.
We devised a similar test for a, which most often means to or at. These
tests did not prove entirely satisfactory, and the latest version of the
program makes no use of them, since where the different aspects of
a homograph belong to different word-classes we now have grammars
precise enough to make the necessary distinction, while homographs
possessing two or more aspects belonging to the same word-class are
sufficiently rare to permit the outputting of alternative translations from
which the reader can choose the correct one.

Macro-semantic tests have not yet been incorporated into the system.
They involve marking in the dictionary words which belong to speci-
cific semantic fields, such as bio-chemistry, chess, or economics, and
keeping a count during the translation process of how many such
words from each field have occurred in the text. One could if necessary
establish the subject of the text in this way by conducting a preliminary
scan of the vocabulary before translation commenced. Any homograph
which had as one of its possible translations a word in English belong-
ing to the semantic field of the text would be translated by that word.
For example, if cavalla occurred in a text which the macro-semantic
test revealed to contain many words referring to chess, it would be
translated as knight; otherwise as horse.

The next stage, which in the most recent version of the program
follows immediately upon the dictionary search, is the application of
the Italian grammar. This is an analytical grammar, working from the
surface or lexical level of the sentence towards the deeper underlying
structures, and is based upon familiar Immediate Constituent Analysis
principles, with labelling of elements. Originally we did not attempt
to produce an absolutely strict grammar, since our object was not
to exclude ungrammatical sequences (there should, after all, have been
none in the input!), but to assign structures to grammatical ones. Thus
all nouns, irrespective of number and gender, were assigned to a single
word-class, and all adjectives, regardless of agreement in the ending,
to another. We had not realised at the time how useful a really strict
grammar could be in the resolution of homographs whose alternative
translations do not belong to the same word-class. For example, stato
is a very common word in Italian, and it is generally quite impracticable
to try to establish by purely semantic methods whether it should be
translated as been or as state; but a precise grammar of Italian such as
we now envisage would certainly reject as ungrammatical a sequence
containing the symbol for “masculine noun beginning with impure S”
where it should contain the symbol for “past participle masculine
singular”; or, of course, vice versa. The computer can therefore ring 
the changes on the grammatical symbols which the dictionary gives 
as possible representations of each homograph, and output a translation 
based upon the sequence which is most acceptable to the grammar.

When the constituent structure of the sentence has been determined 
by the grammar, the next step is the application of the Substitution 
List. This is a roster of all structures and fixed idioms occurring in 
the source language but not acceptable in the target language, together 
with an indication as to what structure or idiom in the latter would 
be both acceptable and semantically equivalent. For instance, the se-
quence NOUN - ADJECTIVE, where the adjective refers to the noun, 
is very common in Italian (for example, bambino intelligente), but scarcely 
ever heard in English except in a few set phrases such as courts martial. 
Therefore the Substitution List specifies this as an unacceptable struc-
ture, to be replaced by ADJECTIVE - NOUN. Thus bambino inte-
ligente will be translated as intelligent boy. Similarly with fixed idioms: 
the lexical sequence sempre più is listed as being unacceptable if translated 
literally, and is replaced by more and more. The application of the Sub-
stitution List is the last stage in the translation process, and the resulting 
English version can now be output.

The results of this initial research suggest that a practical MT sys-
tem based on these methods is a distinct possibility. Furthermore, 
it need not necessarily be limited to Italian as the source language. 
Given adequate grammars and dictionaries, there seems no reason why 
any language capable of transliteration into the Roman alphabet on 
a strict symbol-for-symbol basis should not be used as a source language 
within the system. In the matter of target languages we are much 
more restricted. The system was devised specifically for the purpose 
of translation into English, and while one might substitute for English 
some equally analytic language - for example, Afrikaans - translation 
into a more synthetic language such as French or Russian would prob-
hably present such difficulties as to make it more profitable to develop 
a new system for the purpose. The dictionary look-up and source 
language grammar application would be common to both systems, 
so the work involved would not be as great as it was for devising the 
original system. For the time being, however, research at Cardiff will 
probably continue to concentrate on English as the target language.

(A section of the following translation from Russian to English, 
produced by the Cardiff MT system in July 1973, was read to the 
audience).
АКАДЕМИЯ НАУК СССР
ИНСТИТУТ РУССКОГО ЯЗЫКА

Обзор работ
по современному русскому литературному языку
за 1966 - 1969 гг.

РУССКИЙ ЯЗЫК В ИССЛЕДОВАНИЯХ
ПО АВТОМАТИЧЕСКОМУ ПЕРЕВОДУ.

Под редакцией
члена-корреспондента АН СССР
Ф.П. Филина

(Материалы для обсуждения)

Москва
1973
Редактор выпуска С. К. Шаумян
Авторы: Ю. Д. Апресян (часть ІІ), И.А. Мельчук (часть І)
ОГЛАВЛЕНИЕ

Предисловие ........................................... 5
Список литературы ...................................... 8
Часть I .......................................................... 25
  I. Морфология ........................................... 21
  II. Синтаксис .............................................. 23
    1. Представление синтаксической структуры ...... 24
    2. Обнаружение синтаксической структуры ...... 27
  III. Семантика ............................................. 39
  IV. Русский язык в действующих системах АП ...... 50
Часть II .......................................................... 56
  1. Правильная синтаксическая структура .......... 56
  2. Синтаксическая омонимия ............................ 64
  3. Заключение .............................................. 69
ПРЕДИСЛОВИЕ

В последние годы все увеличивается поток работ, посвященных современному русскому литературному языку. Руководитель уже не может внимательно следить за всей этой литературой. Фонетист часто не знает новых достижений словообразовательной теории, специалист по словообразованию передко не владеет новыми идеями в области синтаксиса и т. д.
FORTRAN IV PROGRAM BABEL (COMPOSED AS UBGR0205F01P) STARTED
30/07/73. TIME 19:08:39

ACADEMY OF SCIENCES U.S.S.R.
INSTITUTE OF RUSSIAN TONGUE.
SURVEY OF WORKS ABOUT CONTEMPORARY RUSSIAN LITERARY
TONGUE OVER 1956 - 1969 (YEARS).
RUSSIAN TONGUE IN INVESTIGATIONS ABOUT AUTOMATIC
TRANSLATION.
UNDER EDITORSHIP OF CORRESPONDING MEMBER AN U.S.S.R.
OF F.P. FILIN.
EDITOR OF ISSUE IS S.K. SHAPMYAN.
AUTHORS: YU.D. APREYAN (PART II.), I.A.
MYEL'CHUK (PART I.).

CONTENTS.
FOREWORD S.
LIST OF LITERATURE S.
PART I. 15.
I. MORPHOLOGY 21.
II. SYNTAX 23.
1. (RE)PRESENTATION/IDEA OF SYNTACTIC STRUCTURE
24.
2. REVELATION OF SYNTACTIC STRUCTURE 27.
III. SEMANTICS 39.
IV. RUSSIAN TONGUE IN WORKING SYSTEMS AP 50.
PART II. 45.
1. RIGHT SYNTACTIC STRUCTURE 56.
2. SYNTACTIC HOMONYMY 64.
3. CONCLUSION/IMPRISONMENT 69.
FOREWORD.

IN LAST YEARS ALWAYS INCREASES STREAM OF WORKS, DEVOTED TO
CONTEMPORARY RUSSIAN LITERARY TONGUE. SPECIALIST IN RUSSIAN
ALREADY CANNOT ATTENTIVELY TO WATCH OVER ALL THIS LITERATURE.
PHONETICIAN OFTEN DOES NOT KNOW OF NEW ACHIEVEMENTS OF WORD-
BUILDING THEORY. SPECIALIST ABOUT WORD-BUILDING OFTEN DOES NOT
POSSESS NEW IDEAS IN PROVINCE OF SYNTAX AND SO ON.
REFERENCES


G. Delconte, Problemi e Metodo, in «Delta», IX (Sept. 1968).
