The Utilisation of the Structural Type-Based Routines as a Homograph Resolution Method in the AMPAR System

A. N. KISELEV

All-Union Centre
for Translation of Scientific
and Technical Literature
and Documentation,
ill. Krzhizhanovskogo, 14,
117218, Moscow, USSR

The special theory of determination of categorial meanings of homographs is used as the basis for an algorithm for analysing English homographs without any ending or ending in -s. The technique for devising homograph resolution routines is described. A description is given of the experiment aimed at determining the effectiveness with which the suggested algorithm can function.

The homograph resolution algorithm developed in the All-Union Centre for Translation of Scientific and Technical Literature and Documentation provides for analysis of English homographs without any ending or ending in *-s.* The forms ending in *-ing* and *-ed,* were involved only when checks run for those endings produced a reliable result and helped resolve homography or narrow it.

What constitutes the basis of this work is the supposition that we need to create, within the framework of working theories like the machine translation by correspondences (MTC) model [1] simulating translation proper for a given language pair as a whole, certain special theories that describe the approach to resolving particular types of ambiguity, and which are based on the contextual determination method.

These special theories should lead to identification of diagnosticating structures (structural types) determining the choice of this or that meaning in resolving particular ambiguities. The diagnosticating structures are in essence correspondences in which the interaction of various levels of language in resolving this or that type of ambiguity is reflected. Identification of such correspondences is a unique task whose solution ensures the optimal ordering of language facts in machine translation.

Two components are usually identified in the contextual determination theory:

1. The type of text perception, i. e. the context is perceived not only as a certain sequence of symbols with certain dictionary information assigned to each of them, but also as a hierarchically organised combination of information levels. Those levels overlap, the task being to separate them by analysing the elements identifying those levels.

2. Typology of contextual dependences, i. e. the differential approach to the contextual dependence of each unit, identification of units whose meanings depend completely on the context, and also context-independent units,

As far as grammar is concerned, it is not expedient to consider within its framework the dependence of each unit on a context although for some units this has to be done. The essence of grammar rules is that they describe a phenomenon for a whole class of words. Therefore, the non-homogeneity as a function of the context should be assessed differentially not only for each unit, but also for classes of words singled out in the process of solving this or that ambiguity.

Thus, a distinctive feature of the approach to the typology of contextual dependence lies in the concreteness expressed in the fact that the type of contextual dependence, as represented by the given structural type in which the ordering of information levels takes place, is identified with due consideration being given to a particular type of ambiguity.

Let us describe the special theory of contextual de~ termination of the categorial meanings of homographs. The 'determinant' and the 'structural type' are the central concepts of this theory.

The determinant is a lexical-grammatical class or subclass, a word list or a single word, a morphological marker, or a definite homograph, or a certain homography class which, singly or in a certain combination, determines one grammatical meaning of the homograph in the language under consideration.

In identifying the categorial meanings of homographs, the following determinants are used:

1. The lexical-grammatical (LG-determinant), which is a lexical-grammatical class or subclass of words having a common semantic content and the general characteristics of distribution. Considered here are the traditional parts of speech — noun, verb, etc., and also subclasses of words such as animate nouns, noun-pronouns, etc. Thus, if *to* is followed by a verb without any ending, it is resolved as an adverb (the particle 'to' is classified as an adverb).

2. The lexical determinant (L-determinant), which is a single word or a list of words. For example, if the definite article *the* is immediately followed by the homograph *only,* the latter is identified as an adjective.

3. The morphemic determinant (M-determinant), which is a suffix or ending. For example, if the homograph *civilian* ends in *-s,* we conclude that it is a noun.

4. The OM-determinant, which is a complex determinant reflecting simultaneously the check for the LG-determinant and also a certain class of homography, or only for a certain class of homography. For example, *excepting*+noun (or a homograph of any class comprising a noun). In this case the word *excepting* is identified as a preposition.

5. The K-determinant, which is a particular word in a certain homography class, differing in its distribution respect from other words of the given class. The particularisation of the word *only* in the 'adjective-adverb' homography class can be an example of the K-determinant. If *only* is immediately preceded by the definite article *the,* it is determined as an adjective, for example, *the only child of his parents.* Other words of the same homography class, for example, *alone* and *sure* do not possess this distinction.

The 'structural type' is the following kind of construction:

T (1,2......n) → X(l,2,...,6)

where T(1,2 ....n) — generalised determinant of the first or second order; X(l,2...,6)—generalised categorial meaning of the homograph of the six possible.

Quantitatively, the structural type indicates that in the resolution routine of a certain type of homographs, there are always n determinants of a certain kind to each of which one of the six categorial meanings corresponds in accordance with the type of homography.

Qualitatively, the set of determinants in the structural types makes it possible to determine the degree of their complexity and arrange them for the purpose of optimal algorithmisation.

For separation of homographs, in the present study we have used the semantic frequency dictionary compiled on the basis of texts of English-language articles in the fields of electronics and computer technology and their Russian translations from the collection of the All-Union Centre for Translation, as well as other sources of about 300,000 word usages for each language [2].

The dictionary entry may be either a word-form or a word combination, whose correct word-for-word translation is impossible, i. e. set expressions.

Indicated for each dictionary entry are: its rank, frequency, the number of texts in which it occurs, and, most importantly, all Russian equivalents used in translating the initial corpus of the English texts.

One of the tasks of the researcher was to explicitly identify homographs on the basis of the semantic frequency dictionary of homographs by studying translation equivalents of English units —the ability of this or that English unit to have translation equivalents belonging to various parts of speech. A complete scan of dictionary entries has been made and, in particular, of all translation equivalents of the English units in the dictionary in order to identify what parts of speech they can belong to. For distributing homographs between homography classes, the following parts of speech were used, as provided for in the system: verb, noun, adjective, adverb, preposition, and conjunction.

As a result of the analysis thus made, 25 homography classes have been identified, for example, verb-noun, noun-adjective, verb-noun-adjective, etc, The described principle of dividing homographs into classes provides the basis for creating tables of correspondences between homographs and particular parts of speech into which they can fall in real usage. The tables of homography contain all homographs in the source English dictionary. A separate table has been constructed for each class of homographs.

At the morphological level, the main means of homograph resolution is the comparison of information on the stem of a word with information on its ending. For example, in the English language the *-ing* and *-ed* endings are characteristic of verbs, the -s ending can belong to the noun and the verb, but not to the adjective. If the homonymic stem *present* has ending *-s,* the comparison of the information for the morphemes which are part of the given word form, while failing to resolve homonymy completely, still, in any case excludes an adjective as a possible part of speech, i. e. narrows the homonymy to the 'verb-noun'.

All homographs with *-ing* and *-ed* endings are determined as verbs. This is done to enable all such words that are assigned the functions of verbal noun, participle, or verbal adverb, to be formed from the verbal stem of the Russian translation equivalent at the stage of synthesis.

The -s ending resolves homography completely for seven homography classes and narrows homography to 'verb-noun' in five cases.

On the morphological level, the English language is patently inadequate for a satisfactory and complete resolution of homography. What is needed is a transition to higher levels — syntactic and semantic. This is effected with the aid of routines analysing the contextual environment of homographs by means of searching the foregoing determinants.

The procedure involved in creating a homograph resolution routine consists of several stages:

,1. Designing a preliminary homography resolution routine.

2. Determination of the place of the routine in the stage,

3. Determination of the general output of the routine.

4. Simplification of the routine.

5. Coding the routine in standard operators, and inputting it into the computer.

Let us consider each of the foregoing stages.

Designing a preliminary homography resolution routine is effected for each homography class by successive identification of the determining contexts. In this process, the concordance compiled for the frequency dictionary is used for selecting all homographs belonging to any one class. Their determining context, fixed as a branch of the routine with 'yes' and 'no' replies, is identified. First the immediate environment of the given homograph is analysed according to the concordance, then this analysis is extended to the entire sentence. This allows identification of both the general characteristics of usage of homographs of a given class, and also of individual resolving contexts for particular homographs of this class. In analysing individual homographs, discontinuous structures having subordinate clauses (for example, in resolving the homography of the word as) are 'taken into account. It is not necessary to transcend the boundaries of the sentence in solving this problem.

For facilitating and accelerating the search for determinants forming the structural types, irrelevant elements were omitted in accordance with the special rules [3].

The regularities of normative English grammar were relied upon in the process of analysing texts by concordance in searching for determinants and identifying structural types. Thus it is obvious that if an adjective is followed by a homograph of the Verb-noun' class, the latter is identified as noun. If a modal verb precedes a homograph of the same class, it is possible to identify it as a verb, etc.

Designing the preliminary routine for resolving homography of any one class corresponded to the requirements of the systemic organisation of the dynamic MTC model component. In particular, provision has been made for expanding the designed routines and making changes in them.

The determination of the place of the routine in the stage. An important question that arises in devising an algorithm for resolving homography is the determination of the place the routine for that particular homography occupies within a stage, or, in other words, the sequence of routines within the stage, which distinguishes the given approach from all kinds of syntactic analysis where the sentence structure is represented as a tree. The given approach is based on the following criteria determining the sequence of routines:

a) the criterion of simplicity, which means that, in the first place, simple routines are identified in which non-homographs are used as determining features: those routines are also simple in terms or their volume;

b) the criterion of frequency, which means that routines for homography resolution operate in descending order of frequency of their occurrence because resolution of high-frequency types of homography provides a larger volume of information and helps resolve a lower-frequency homography.

Determination of the general output of the routine. All categorial meanings of homographs are, if possible, determined in the homograph resolution algorithm of the MTC model. In cases when determination is difficult or impossible, determination of the general categorial meaning of a specific class of homography or the general output of the routine can be relied upon. In this approach, the principle of structural types diversity has been employed in selecting the main or general categorial meaning of an homograph. Research on determining the diversity of structural types of any one categorial meaning of the homograph by classes thereof has been done in an automated fashion. First, on the basis of the structural types identified with the aid of the concordance and fixed as branches of the preliminary routine, the routine has been programmed in standard operators [4], with the choice of the general output of the routine being made arbitrarily. Then, based on an analysis of the routings* followed in resolving the homography, resulting from the routine operation for real texts and produced in the automated fashion by the machine, the meaning with a wider diversity of structural types was selected as the general output of the routine.

Simplification of the routine is achieved by discarding those structural types which determine the meaning coinciding with that of the routine's general output. Fir|t and foremost, complex structural types having many lift and right checks and wide omissions undergo such simplification or discarding. In the process of discarding, one should not forget, however, that a portion of the structural types determining the general categorial meaning of the homograph should be retained in the routine for convenience of the logical construction and acceleration of its functioning, because the given branch may turn out to be frequent.

Coding of the routine in standard operators and inputting it into the computer. After all preceding stages are completed, that of coding the routine in standard operators follows. The standard operators are a specialised programming language that permits linguists to directly participate in devising and correcting the MT routines.

On the basis of the technique described, a routine is devised for each of the identified classes of homography with its own set of structural types that permits the part of speech of any homograph in the text to be determined and makes possible further syntactic analysis. The routine functions in the following way. Any homograph occurring in the text is taken as a current word. Then, analysis of the contextual environment of the homograph begins by checking whether it belongs to some structural type of the routine. If this is the case, the part of speech it belongs to is determined. If the homograph does not belong to any of the structural types of the routine, a categorial meaning is assigned to it at the general output.

Experimental research was done to study the determinants used in resolving homography, and to single out structural types identified for each class of homography under consideration and relevant to the determination of some particular categorial meaning of the homograph. Structural types were identified in practice by routing the homograph resolution routines. Research was carried out automatically and manually. The automated part of the research was done with the computer, the routine entered into it and coded in standard operators in the form of a special algorithm, producing a routing through the routine, i. e. a path in the algorithm, along which the resolution of that particular homography type was made. The aim of the manual part of the research was to trace, on the basis of the identified routings, the entire path of the analysis in the routine of a given type of homography, discarding the branches with a 'no' reply and coding the determinants, i. e., each check for 'yes', with the aid of special designations.

As a result of this procedure, sets of structural types characteristic of each class of homography were obtained. A total of 322 structural types were identified, 207 of which were different. Out of the 207 structural types, 98 were with the K-determinant. The number of non-specified structural types was 109. This goes to show that specification plays a big role in resolving an homograph. On the other hand, the predominance of the structural types without the K-determinant points to the fact that the determination of homographic meaning has a generalised nature. This is also supported by the fact that a greater part of structural types with the K-determinant (73) are encountered only in 5 of 28 routines. Among the structural types, there were also those characteristic both of all homograph resolution routines and

* Routing is a certain sequence of steps ('yes' or 'no') which is followed in resolving homography in a particular routine on the basis of a homograph's real use in the text.

of only part of them. This makes it possible to talk about the universality of structural types in general, and in particular, about a limited universality, i. e. specific only for routines of certain types. In other words, the broader the context, the less universal it is in resolving homographies of various classes.

The data obtained in studying the number of determinants in the structural types are given in Table 1 below.

Table 1

Number of determinants in a structural type	Number of structural types with a given number of determinants	Number of determinants in a structural type	Number of structural types with a given number of determinants
1	75	6	7
2	108	7	3
3	70	8	3
4	40	9	1
5	15	10	—

It is evident from Table I that the number of determinants within the structural types seldom exceeds four, which goes to confirm the results obtained by Yu. N. Marchuk [5] in studying polysemantic words. Structural types consisting of two determinants are most frequent. For applied purposes—MT in particular — this conclusion is very important because it points to the adequate simplicity and economic efficiency of the given method of determining the categorial meanings of homographs.

The homograph resolution routines are economical because they do not provide for a complete description of the object, not requiring, in particular, the construction of a complete syntactical structure of the sentence; rather by local analysis of the context (operator method), structural types consisting of a small number of identification features (determinants) are put together.

As a result of the research, it has also been revealed how the determinants are used within the structural types of the homograph resolution routines. The respective data are cited in Table 2 (the selection has been made on the basis of 11 homograph resolution routines).

Table 2

The type of determinant	Number of uses
LG	235
L	88
M	58
OM	57
K	91

As is evident from Table 2, the LG-determinant is used in resolving homography in an overwhelming majority of cases, which shows that the determination of the categorial meanings of homographs at the level of identified classes of words (parts of speech) close to the traditional parts of speech is quite satisfactory. In 10 of 28 homograph resolution routines, specification of the analysed homograph has been used. Homographs, such as *for, over, off, more, still, like, very,* that are peculiar from the viewpoint of their distribution, have been particularised. According to our observations, this helps in determining their correspondence to particular parts of speech.

The OM-determinant is used quite frequently. This points to the existence of a large number of homograph chains in texts, as well as to the fact that, in resolving certain kinds of homography, the utilisation of OM-determinants in combination with other determinants, for example, the LG-determinant, produces good results. All cases of OM-determinant utilisation have been fixed in the structural types.

The assessment of the effectiveness of the homograph ' resolution algorithms (routines) had two purposes: adding new structural types to algorithms for enhancing their resolving power, and identifying the main types of errors that permit assessment of the capabilities of the method.

The homograph resolution algorithms were repeatedly verified in the course of their elaboration, and also in the course of their pilot and industrial utilisation in translating texts in the field of computer technology and programming.

The main types of errors have been identified on the basis of intermediate printouts of translated texts. First, the text is printed out with indication of information cells in the state immediately preceding, the stage of homograph resolution, and also with indication of homographs and a homography class. Then, after running the homograph resolution stage, the text was printed out again. In this variant of the text, all homographs were already assigned particular parts of speech. After that, the cell-by-cell printouts obtained before the homograph resolution stage and the cell-by-cell printouts obtained after the run of the stage were compared with each other.

To identify the types of errors, texts containing 100,000 word usages in total volume were analysed. 2,900 of 30,000 homographs, i. e. about 10 per cent of all homographs, were resolved incorrectly. The results of the analysis of incorrectly resolved homography are given in Table 3.

Thus, by expanding the concordance, up to 85 per cent of errors can be eliminated. On average, for homographs of all classes, the described method produces 92 per cent of correctly resolved cases. A majority of errors is connected with the 'verb-noun' class of homography. At the present time, an analysis is being made of the results of homography resolution with the use of stable files of texts in the given sublanguage of computer technology. It is planned to carry out the same analysis in the field of aviation.

The research has permitted the following conclusions to be drawn:

I. The algorithmic modelling of the translation process, and, in particular, the creation of the intermediate MTC model in which the specific features of language comparison in the process of translation find their reflection calls for a strict and precise approach to terms, such as 'homonymy', 'ambiguity', and 'homography'. By the term 'homography' one should understand the ability of the word-form with zero ending or *-s* ending to have several categorial meanings (to belong to various parts of speech) .

Table 3

Type of error	Errors In the examined volume. In %
The determination process in the algorithm is incorrect because the homograph chain contains three or more elements of various types. The error can be corrected with the aid of the given method by adding new structural types.	20
The case of the imperative mood of the verb for the 'verb-noun' class is left out of consideration in the algorithm. In principle, the error can be corrected by means of a wider involvement of punctuation marks as determinants.	8
There is no determinant in the algorithm; a correct establishment of correspondences can be achieved by including the chain into the stage of word combinations processing.	15
The general output of the routine is incorrect.	2
Determination process within the framework of the method is rather difficult or impossible.	15
The isolated homograph or homograph in a chain of two elements are resolved incorrectly: — because of the absence of some structural type in the algorithm which is mostly connected with the necessity to particularise a homograph; — because of an excessively narrow rule of omission in the search for the determinant; — because of technical errors in the routine or the dictionary. Can be corrected.	40

2. The utilisation of the idea of approximative linguistic computation makes it possible to quite effectively organise the search of multi-level translation correspondences on the basis of preliminary homograph resolution by simulating a human translator's actions to a certain extent in terms close to the traditional ones.

3. The establishment of correspondences can be modelled in a satisfactory fashion on the basis of the special theory of contextual determination of the categorial meanings of homographs and by way of identifying structural types relative to homograph classes.

4. Structural types of various degrees of complexity identified in determining the categorial meanings of homographs in diverse combinations for unlike classes, consist of a wide finite set of determinants and can solve both the frequent and rather rare cases of homography with priority given to solving the most frequent, mass cases thereof. Structural types equivalent to each other as well as types characteristic of all homograph resolution routines have been identified. The identified structural types permit accurate description of the contextual dependences both quantitatively and qualitatively for a wide language field. The created special theory of determination of the categorial meanings of homographs allows us to estimate the effectiveness of this method and list the cases requiring more subtle methods of analysis.

Identification of the structural types has made it possible to perceive two tendencies in utilising the determinants for homograph resolution. First, the utilisation of the specification technique (the K-determinant) with regard to certain classes of homography. Second, the possibility of going from the L-determinant to the LG-determinant with regard to a number of homography classes, for example, with regard to the 'verb-noun' class, which can be achieved by identifying, on the basis of distributive method, and generalising lists of semantic subclasses of the given LG-classes.

The research results are important for a further investigation of the phenomenon of ambiguity and classification of types of ambiguity, and show a way of solving this problem. The special theory of determination of categorial meanings of homographs and the structural types identified on its basis are part of the transitional working MTC model. The orientation of the MTC model towards translation, in combination with the obtained efficiency of the given homograph analysis method, indicate that the given method of homography analysis or a method equivalent to it is a necessary stage of work.

A higher efficiency of MT in general and of the stage of homograph resolution, in particular, can be achieved on the basis of the orientation of homograph resolution routines to the fields of discourse. The texts of invention descriptions, abstracts, journal articles, and specifications, as well as headings of articles, are specific both lexically and grammatically. And, consequently, the given variant of the MT system will depend on the nature of the text. This fact, in turn, means that there will be both quantitative and qualitative differences of structural types in homograph resolution routines.

The foregoing method of homograph resolution based on the MTC model is suitable both for practical utilisation and for a further development of the MT theory.

REFERENCES

1. Marchuk, Yu. N. *The problems of machine translation.* Moscow: Nauka Publishers, 1983. 232 p.

2. Ubin, I. I. The utilisation of semantic frequency dictionaries and concordances in compiling machine dictionaries. In: *The machine translation and automation of information processes.* Moscow: All-Union Centre for Translation, 1975, 98—108.

3. Marchuk, Yu. N.; Tikhomirov, B. D.; Shcherbinin, V. I., A system of machine translation from English into Russian. In: *The machine translation and automation of information processes.* Moscow: Ail-Union Centre for Translation, 1975, 18—33.

4. Tikhomirov, B. D.; Drambian, L. V.; Istomina, I. M. et al. The specialised language for programming machine translation algorithms. In: *Theses of the reports delivered at the International Seminar for Machine Translation.* Moscow: Ail-Union Centre for Translation, 1979, 60—62.

5. Marchuk, Yu. N. The experience of machine realisation of the distributive method for determination of lexical meanings. In: *Statistics of speech and automated text analysis.* Leningrad: Nauka Publishers, 1973, 181—230.

[*All references are in Russian*]