Syntactical Variants†

Bjarne Ulvestad, Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge, Massachusetts*

Traditional grammar is normally eclectic and vaguely formulated, and it often tends to overgeneralize or fails to state the range of validity for its rules. Grammars for mechanical translation must be all-inclusive and rigorously explicit. While the input language grammar must register all the grammatical constructions possible, the existence of basically synonymous morphological and syntactical variants permits considerable inventorial reduction in the output grammar. These considerations are discussed with reference to English and German examples: verb phrases with 'remember'/ (sich) erinnern as the head; 'as if' / als ob clauses.

IT IS POSSIBLE to imagine a series of poor but successively 'better' machine-made translations, ranging from, say, 'very poor' to 'fair' or 'not so very poor,' which might be found to be substantially adequate for their various purposes. Thus even a lowest-grade or 'very poor' translation would conceivably have a demonstrable adequacy, provided its purpose were merely to acquaint its prospective readers with the subject matter of the original (input language) text.[1] Leading up from this kind of primitive, low-standard mechanical translation to one that would be regarded by the pundits as 'correct,' to the finest shades of idiomatic nuances, there is an almost discouragingly long, devious path, or rather a long series of shorter excursions each of which is more complex and laborious than its predecessor. If we, as we should, consider it imperative never to compromise with perfection where perfection is attainable, all the words and all

the syntactical constructions of a given pair of languages, and especially of the one on the input side of the translation machine, will ultimately have been 'tagged' or assigned their specific memberships in a large number of groups and subgroups of linguistic entities, and the more exhaustive this intricate taxonomy, the more adequate, i.e., the less liable to produce ungrammatical and nonsensical sentence sequences, will be the corresponding translation mechanism.

The tantalizing question as to whether an absolutely foolproof apparatus for the mechanical transfer of information from one language to another can be constructed, if only in theory, need not bother us too much at this stage, for even if the answer to the question should in the end turn out to be negative, less-than-perfect mechanical translation will nevertheless be useful for scholars, whose main concern is naturally to obtain an adequate communication of scientific facts and ideas rather than stylistically impeccable texts, desirable though the latter may be.

Judging from reports on the highly significant work which is at present carried on at various universities, we have every reason to believe that most of the general technical problems of mechanical translation are approaching their solution. As an example of this kind of promising study, one may mention N. Chomsky's and V. Yngve's research into workable recognition devices for use in sentence-for-sentence translation, which is vastly preferable to word-for-word transfer. While the bulk of linguistic work in the field of mechanical translation has thus far admittedly been of a rather general

1. Cf. J. W. Perry, "Translation of Russian technical literature by machine," MT, Vol. 2, No. 1, pp. 15-24 (1955).

and preliminary nature, researchers on both sides of the Atlantic are becoming more and more aware that the most pressing requirement for further progress is the composition of total-coverage grammars deliberately executed with mechanical translation in mind. We do not have such grammars for any language, except in rudimentary and fragmentary form, but even at this early date we can discuss some of their conspicuous features, as distinct from those of what we may term traditional grammars.

In this article a few problems in mechanical translation grammar will be presented and discussed, with some reference to their practical relevance to the input language and to the output language. English and German are the two languages chosen for this exposition. However, substantially similar problems will no doubt be found in any language.

We can state without reservation that in constructing grammars for the input language and for the output language, the input grammar must be subjected to the more piecemeal examination of particular problems. One of the most transparent reasons for this lies in the relatively large number of basically isosemantic morphological and syntactical variants that exist in every linguistic system. While all these variants will presumably have to be identified and registered in the input language grammar, considerable reduction in the number of corresponding variants will ordinarily be possible in the output grammar, as will be seen below. It must be emphasized that the chief difference between traditional grammar and what may be called mechanical translation (input language) grammar is that the former is eclectic and normally vaguely formulated, whereas the latter will be all-inclusive and rigorously explicit and formalized. Traditional grammars overgeneralize and rarely state the actual range of the validity of each rule; mechanical translation grammar must, ideally, explicate all the cases for which the given rule applies as well as those for which it does not. Furthermore, mechanical translation grammar must of necessity account for the total number of linguistic constructions that occur in a given language even if traditional grammars categorically state the nonoccurrence of certain members; [2] and misleading transformation rules must be recognized as such and correctly restated. [3] Whereas variant constructions of low statistical probabilities may on the whole be disregarded in the grammar of the output language, [4] they cannot, as a rule, be left out of the grammar of the input language without more or less serious consequences for the quality of the eventual translation. It is obvious from the remarks made above that the mechanical translation point of view will compel linguists to examine in detail problems that have hitherto been regarded as trivial or inconsequential. We can therefore expect that mechanical translation research will be of fundamental value to structural linguistics.

The important task of registering all syntactical variants, including those that are ordinarily overlooked in standard grammars, need not necessarily lead to a correspondingly greater complexity on the part of the eventual encoding program, although it may seem so at first glance. An example will perhaps help.

(1) Ich erinnere mich an ihn (den Mann)

(2) Ich erinnere mich auf ihn (den Mann)

(3) Ich erinnere mir ihn (den Mann)

(4) Ich erinnere mich ihn (den Mann)

(5) Ich erinnere ihn (den Mann)

(6) Ich erinnere mich seiner (des Mannes)

These German sentences are built around the weak verb (sich) erinnern 'remember' and corresponding to the English sentences 'I remember him' and 'I remember the man.'

---

2. Cf. B. Ulvestad, "Object clauses without dass dependent on negative governing clauses in modern German," Monatshefte, 47.329-38 (1955).

3. A typical instance is furnished by E. E. Cochran, A Practical German Review Grammar. 11th printing (New York, 1947), p. 241: "Note: zu after sagen is dropped in an indirect statement." The example illustrating this dropping of zu is: Er sagte zu mir: "Ich kann es mir nicht leisten," vs. Er sagte mir, er könnte es sich nicht leisten. That this rule is invalid in its present categorical formulation is seen from such sentences as: Er sagte zu Sabine, er werde sie . . . abholen (Brentano), Franz... sagte einmal zu mir, es gebe in jedem Dorf ein oder zwei schwere Taten (Wittich).

4. This consideration will be taken up for separate discussion in a later article.

Only (1) and (6) belong to the generally accepted standard language, and for that particular code the traditional formula, 'sich (acc.) erinnern is followed by a genitive construction or by the preposition an with an accusative construction,' is correctly stated, provided, of course, that one does not take 'followed by' literally. In normal modern German literary prose, however, one may encounter any one of the six types. Now, if we want to register every one of the sentence types with reflexive erinnern in the input code (this excludes 5), we need only add the verb erinnern not only to the class of reflexive verbs with the reflexive pronoun in the accusative case, but also to the class of verbs that may occur with the reflexive pronoun in the dative, and subsequently state, e.g., that the verb erinnern with accusative reflexive may 'govern' the accusative, the genitive, or a prepositional phrase with an or auf followed by an accusative noun phrase (NP). Since these entities will presumably have been registered and classified in some department of the grammar anyway, they do not have to be restated, but only referred to in terms of a defined code signal. This signal will indicate, for instance, that the verb (sich) erinnern belongs with denken in that it 'governs' an an-phrase with the accusative, and with sehen in that it takes an auf-phrase with the accusative.

If the purpose of the mechanical translation grammar and translation apparatus were restricted exclusively to the transfer of German scientific texts, sentence types (1) and (6) above would probably be the only ones that would need to be encoded. Even for translation of current novelistic prose we need only add (5), which occurs much more frequently than (2) and (3). In this kind of literary prose, the frequency continuum runs as follows, from very high to very low: (6)— (1)— (5) — (2) — (3)— (4).[5] If, on the other hand, a speaker of the Hamburg Umgangssprache were to be used as 'informant,' the first part of the frequency sequence would probably be (5) — (1); (6) can hardly be said to belong in this city language at all.[6]

Whatever the tasks for which the translation machine is designed, the encoding will not be made too difficult by the requirement of full coverage. It is the patient grammar writer whose difficulties are enhanced by new decisions to improve the translation.

It is interesting that if German were the output language, the situation in the examples above would be reversed and considerably less complex. As input, we would have English sentences with the verbs 'remember,' 'recall,' and possibly 'recollect,' all of which are closely related from the point of view of multiple-class memberships. With German as the output language, one of the six types above is sufficient for mechanical translation purposes since we are primarily interested in cognitive meaning transfer, not in the kind of additional information 'natural language' may furnish (age, sex, dialect, education, business background, etc.)

Naturally, the reduction of the number of variants in the output language to one is advisable only if the variants are absolutely free or if there is no possibility of making a meaningful selection out of two or more output variants on the basis of clues found in the input language. We snail explain this below with reference to a typical mechanical translation problem, using as examples German and English clauses which may be termed 'quasi clauses' (in English, 'as if'-clauses; in German, als ob-Sätze). Presentation of a grammar of these clauses for mechanical translation is the purpose of the remainder of this paper.

Variations on the following statement, with its examples, are current in textbooks of German: 'The secondary subjunctive (past subjunctive) is usual after als ob 'as if.' Er sprach, als ob er das Buch gefunden hätte. . . . ob may be omitted and inverted order used. ... Er sprach, als hätte er das Buch gefunden.'[7] It is not difficult to see that this 'quasi clause grammar' is far

---

5. The data for this were obtained from a corpus of 52 recent German novels; (3) and (4) occurred only five and three times, respectively, and there was a considerable frequency drop between (6), (1), and the rest.

6. Native informants refer to (6) as "stilted," "constructed," "archaic."

7. P.H. Curts, Basic German, revised ed. (New York, 1946), p. 71. It does not matter much whether one's description of als (ob, wenn) reads, (1) 'the ob, like the wenn, may be omitted,' or (2) 'the quasi conjunction is als, but ob or wenn may be added,' although logically (1) is preferable in a grammar of the spoken standard (Hochsprache popularly also called Schriftsprache). and (2) better corresponds to the usage actually found in the written (novelistic) language.

too fragmentary to be used except for introducing the 'rudiments of elementary German' to beginners; so we shall not take time to demonstrate its shortcomings. Rather, we shall attempt to write as complete a grammar of the German 'quasi clauses' as possible from the data available to us. Subsequently some practical problems with reference to the transfer processing will be discussed.

Let us consider the following six sentences.

(7) Ihm war, als habe er sie seufzen gehört (Waggerl)

(8) Es war, als ob noch einmal die Sonne, Wasser und Wind ... dem Oberleutnant in dieser Gestalt vor die Augen treten wollten (Tügel)

(9) Mister Wenner ging durch das Dorf, als wenn es gar keine Schwalbacher gäbe (Kirschweng)

(10) Und doch war es, wie wenn ein schieferblanker, tödlicher Ernst sich auf den ganzen Platz gelegt hätte (Goes)

(11) Wenn ich im Fahren lange hinaufsah, war es mir, der ganze Himmel käme auf mich zu (Bauer)

(12) Ich lief schnell, wie als gälte es, sich ein Landgut zu erobern auf diesem Gang (Goes)

Sentences (7) to (12) have different 'quasi' conjunctions (QC's), namely, als, als ob, als wenn, wie wenn, zero (Ø), and wie als. The internal relationships between these sentences will be seen from the following regrouping of (7) to (12) symbolized in terms of significant constituents (the symbol / is read 'or'):[8]

(7) --------, als + Vfin + NP + (Vinf / Vpp)

(12) -------- , wie als----------------------------------

(8) -------- , als ob + NP + (Vinf / Vpp) + Vfin

(9) -------- , als wenn --------------------------

(10) -------- , wie wenn --------------------------------

(11) -------- , Ø + NP + VP ------------------------------

---

8. The mode of the finite verb in the ' quasi' clause is not considered at this point. Note that the term 'Vfin' in parentheses is used in a wide sense and includes so-called passive infinitives such as gehört werden, gehört worden sein, etc.

We symbolize the noun phrase and the potentially succeeding infinitive or past participle under one sign, Z [NP + (Vinf /Vpp) $=$ Z]; and the relationship between (7), (12) on the one hand, and (8), (9), (10) on the other will be seen to be one of constituency permutation to the right of the QC. For further simplification of the structural statements, we may operate with three classes of QC's: QC_1 (als, wie als), QC_2 (als ob, als wenn, wie wenn), and QC_3 (zero).[9] Note that a comma always separates a clause from a succeeding dependent clause and accordingly stands in an immediate concatenation relationship with the conjunction. We can therefore (and this may be useful for mechanical translation encoding) subsume under the term 'conjunction,' for maximum mechanical translation signal power, the conjunction itself with the preceding comma, so that, for example, the symbol QC_1 shall be henceforth taken to mean 'comma followed by QC_1.' The six 'quasi' sentences can accordingly be written as follows:

I. (7), (12)   ----------QC_1 + Vfin + Z

II. (8). (9), (10)   ---------QC_2 + Z + Vfin

III. (11)   ----------QC_3 + NP + VP

Further reduction, stating the transformation relationship between I and II in formal terms, is possible. For instance, one might state the rules: 'for transforming I into II. rewrite QC_1 as QC_2 reversing the order of Vfin + Z, and for transforming II into I, rewrite QC_2 as QC_1 reversing the order of Z and Vfin,' but further study would disclose that T I $\rightarrow$ II is correctly stated, and not the reverse T II$\rightarrow$ I. From er tat, als hätte er ihn nicht gesehen (I) we clearly obtain by this transformation: er tat, als ob er ihn nicht gesehen hätte (II), but there exist instances of so-called elliptic II-sentences that do not permit a direct transformation T II $\rightarrow$ I, for instance, er tat als ob er ihn nicht gesehen, in which the finite verb (here,

---

9. On a different level of analysis, one might make use of the structural relationships between (12) and a sentence such as es war mehr so, als hielte sich etwas an ihrem Bein fest (Nossack) and state that the adverb so in the governing clause can be shifted into the dependent clause and changing its status into that of a corresponding conjunction particle, thus: X + so, als + Y $\rightarrow$ X, wie als + Y. Note the positions of the comma in the two formulas.

hätte or habe) is dropped, or more correctly stated, does not occur. The ellipsis of the (readily predictable) finite verbs haben and sein after past participles is encountered occasionally in all subtypes of II, in (8) as well äs in (9) and (10), whereas the finite verb must always be made explicit in I. And the omission of haben / sein is not restricted to 'quasi' clauses. [Cf. the dependent clauses of sentences like er fragte, ob er ihn gesehen [habe / hätte] and als er nach Hause gekommen [war], fand er, dass. ....] This 'dropping' of haben / sein after past participles thus need not be specially explicated in the grammar of 'quasi' clauses; it will have been taken into account elsewhere. Another distinctive feature differentiating I and II may be adduced: The subjunctive mode of the finite verb, or rather the subjunctive ([er] höre, [er] ginge) or the nonovert, 'neutral, ambiguous' mode (indicative or subjunctive, such as [er] hörte, [er] suchte) is obligatory in the I-sentences, but not in the II-sentences; for instance, er tut, als höre / hörte er nichts, but er tut, als ob er nichts hört / höre / hörte, where hört is an overtly indicative weak verb. In a recent study of German 'quasi' sentences, based on twenty-four novels, no overt indicative finite verbs were found among 737 als-clause s (I), but fifteen were found among the 187 als ob- / als wenn-clauses (II) found in the corpus. [10] Consequently, the establishment of groups I, II, and III appears so far to be the simplest possible classification and if we include reference to the mode of the finite verb in the 'quasi' clause, the following three statements or formulas describe the grammar of the 'quasi' clauses in German:

I. $QC_1 + \text{Vfin subj} + Z$

II. $QC_2 + Z + \text{Vfin subj / ind}$

III. $QC_3 + NP + VP \text{ subj /ind}$

Formulas I and II uniquely define German 'quasi' clauses. They can therefore be used directly, i.e., without additional specification, as clause identification formulas in standard written German. Thus $X + I + Y$ or $X + II + Y$ is normally sufficient information for establishing that one is concerned with sentences or sentence sequences that include

'quasi' clauses, e.g., er sagte, als hätte er nichts verstanden, dass er es morgen Versucher werde. [11] Here the 'quasi' clause is included in an indirect discourse sentence, and its special formula is simply $X + QC_1 + \text{Vfin subj} + Z$. Note that 'Vfin + Z' is an indispensable element in formula I, because of the nonunique function of als as a dependent clause conjunction (cf. als er nach Hause kam, etc.), whereas in formula II the element ' Z + Vfin' can be considered predictable, and the simplified formula $X + QC_2 + Z$ would perhaps be an adequate statement for a sentence like am nächsten Tage lag er ganz still, als ob er tot wäre. The unique function of als ob as a conjunction makes this reduction possible.

Formula III is more recalcitrant in that its primitive form, $(\text{---------}\emptyset + NP + VP)$ is also the statement of the structure of indirect discourse sentences with zero conjunction; e.g., er sagte, er sei krank. Actually, III formalizes a genuine overlapping or ambiguous sentence type. [Cf. such sentences as mir scheint, dass..........., mir scheint, $\emptyset$..........., and mir scheint, als ob ....] Note that our token sentence (11) above can be translated either as '... it seemed to me as though..' or as '... it seemed to me (that)...,' with only trivial difference in cognitive meaning. There are two possible ways of solving the recognition problem in this case: (1) We can add specifications as to the context of the clause and state that zero is used as a 'quasi' conjunction after governing clauses such as mir ist, es scheint, or (2) we can drop III from our 'quasi' clause formulations altogether and consider it an indirect discourse formula only (the term 'indirect discourse' being used here in its traditional meaning). The second solution seems preferable for the following reasons: The zero

10. B. Ulvestad, "The Structure of the German Quasi Clauses," to be published in Germanic Review (1957).

11. This statement needs to be qualified to exclude some rarely occurring clauses that would seem to correspond to II in its present formulations. The following sequence was found in W.v.Niebelschütz, Verschneite Tiefen, (Berlin, 1940), p. 144: 'Doch wessen das Herz hier gierig ist, weiss niemand; nur ich. Vielleicht weiss es der Ritter auch? Mag sein. Mag es sein, es wäre leichter für mich, als wenn ich's ihm sagen müsste.' The clause starting with als wenn means: 'than if I had to tell it to him.' Such dependent clauses as this are found only after comparatives in the governing clauses, here, leichter.

Table I

	als		als ob		als wenn	
	c.pr.	c.pt.	c.pr.	c.pt.	c.pr.	c.pt.
1		63	1			
2	15	15				
3	63	15		2		
4	1	2		1		
5	3	35	1			
6	16	25	5	11		1
7		1	1			
8	9	5	1	2		
9		11		7		
10	1	8		5		
11	1	31		2		1
12	1	6	4	10		
13			1		11	12
14	1	38				
15	2	5		5		
16	38	7				
17	6	26		11		6
18	9	18				
19	4	24		7		
20	13	5	1	1		
21	3		1			
22	8	5	2	3		
23	10	14	1	4		
24	1	3	2	7		
Sums	205	362	21	78	11	20

Frequencies of chosen present subjunctive (c.pr.) and chosen past subjunctive (c.pt.) in three different 'quasi' clause types in novels by 24 authors.

conjunction occurs only after governing clauses like es scheint, mir ist, es kommt mir vor, and it is infrequently found. Only thirteen examples [such as mir schien, ich könnte sie aussprechen, jedoch fehlte das Wort (Zweig)] were found among 1168 'quasi' sentences taken from twenty-four works. This in conjunction with the basic similarities in meaning ('it seemed to me that / as though ....'), appears to furnish sufficient justification for operating with only two types of 'quasi' clauses, I and II,

and our reduced grammar now simply reads:

I. QC_1 + Vfin subj + Z

II. QC_2 + Z + Vfin subj / ind

The tense-forms of the subjunctive in such clauses need not occupy us for long. In most traditional grammars, which are usually of the prescriptive type, statements indicating the obligatory nature of past subjunctive finite verbs are found. Table I amply demonstrates that these statements are untenable and unwarranted.

---

12. The term 'chosen present/past subjunctive' means that either tense form in a given case would represent the subjunctive mode unambiguously. In other words, we are interested in the ratios between the numbers of occurrence of such forms as, e.g., [er] sei, gehe, bringe (present subjunctive) and [er] wäre, ginge, brächte (past subjunctive). The names of the authors are of no importance in this context.

We would therefore be wrong in adding the word 'past' after 'subj' in formulas I and II; the correct statement is obviously one that does not specify tense-form.   If German were the output language, (in which case we would be faced with a choice, see below) the grammar would read, at least for the literary style level:

  I. QC_1 + Vfin subj past + Z

In this formula, QC1 would include only <u>als</u>, not <u>wie als</u>, and formula II would not occur in this grammar at all, unless compelling reasons for its inclusion were discovered.[13]

A similar problem emerges with regard to the translation of German into English: Should we register both 'as if' and 'as though' as correspondent conjunctions, and if not, which one would be preferable?   Let us discuss this from the point of view of a particular transfer situation.   The following German sentences are all grammatically correct:

> Er tat, als ob er krank wäre
>
> ------ , als wenn----------------
>
> ------ , wie wenn --------------
>
> ------ , als wäre er krank
>
> ------ , wie als ------------

These sentences are, at least from the point of view of mechanical translation, isosemantic and can be translated as either 'he acted as if he were ill,' or 'he acted as though he were ill.' Therefore, NP + VP + 'as if' + NP + VP seems just as good a correspondence formula as NP + VP + 'as though' + NP + VP.[14] However, we would reasonably argue that the slightly 'elevated,' 'literary' connotation of 'as though' in contradistinction to the more 'colloquial' one of 'as if' corresponds to that of the German <u>als</u> (I) and <u>als ob</u> (II), respectively, in which case one may suggest as an

adequate German-to-English transfer grammar of 'quasi' clauses:

  I. QC_1 + Vfin subj + Z
        → 'as though' + NP + VP

  II. QC_2 + Z + Vfin subj / ind
        → 'as if' + NP + VP

The concise 'quasi' clause grammar which we have worked out above could be further simplified within the context of a full-scale input grammar of German, because most, perhaps all, of the constituents would already have been described and classified.   For instance, the two clauses in the sentence <u>wenn</u> er <u>mich sähe,</u> <u>würde</u> er <u>grüssen</u> belong in the same classes as some of the 'quasi' clause constructions after <u>als</u> in [<u>er tat,</u>] <u>als wenn</u> er <u>mich sähe</u> and [<u>er tat,</u>] <u>als würde</u> er <u>grüssen,</u> respectively.

The classification and coding of sentence elements and the subsequent elaboration of the simplest possible grammatical rules in terms of these classes are indispensable preliminaries to a successful construction of a workable translation machine.   Every new grammatical statement will also represent a step forward in our scientific description of the language whose structure the grammar explicates and formalizes.   The ultimate grammar will constitute the central prerequisite for a translation machine.

---

13. The reasons for preferring I (with <u>als</u>) to II (with <u>als ob</u>, <u>als wenn</u>) for the output grammar, if only one formula were to be employed, can be read out of the table.

14. A more complete discussion of the English correspondences would, of course, include such 'quasi' clauses as 'as though being ill.'