

Mechanical determination of the constituents of german substantive compounds

Erwin Reifler, Far Eastern Department, University of Washington, Seattle

The MT process comprises four distinctive sub-processes called the *input*, the *identification* of input forms, the *translation process proper* and the *output*. Initially certain linguistic phenomena seemed likely to prevent the complete mechanization of the identification process. The problem is the following.

Identification presupposes a record of things remembered, with which everything to be identified is compared. An essential feature of all MT systems will be the "mechanical memory" which corresponds to the bi-lingual dictionary plus the knowledge at the disposal of the human translator. The head entries of this memory will consist of individual free and bound forms and idiomatic sequences. All input units whether they be words, portions of words, or groups of words will first have to be identified with their "memory equivalents" before their "output equivalents" can be determined mechanically.

Many important languages include large numbers of compound words which, though they are mostly of low frequency, are essential for understanding the context in which they occur. These compound words are made up of a comparatively small number of constituents, many of which also occur as free forms of higher frequency. German examples of the latter are *Hoch* (high) and *gefühl* (feeling) in *Hochgefühl* (exalted feeling) and *mittag* (noon) in *Nachmittag* (afternoon); *Nach* (after) in *Nachmittag* is an example of a very high frequency constituent.

It is natural to think of economizing coding and access time by excluding large and, in fact, continuously increasing numbers of compounds from the mechanical memory, and adding instead the comparatively few constituents which are *productive*—that is, are found in more than one compound—and do not occur as free forms. An example is German *seitig* (-sided) in *einseitig*, *zweiseitig*, etc., (one-, two-sided, etc.). Constituents which also occur as free forms are entitled to a place in the mechanical memory a priori. Such an arrangement would permit the identifica-

tion of compounds by means of the mechanical identification of their constituents. This would result in a welcome reduction of the size of the mechanical memory. It is true that the matching of each compound would be replaced by the matching of its two or more constituents, and the design of the matching mechanism would have to include provisions for the dissection of compounds into their constituents. Nevertheless, because of the comparatively low frequency of most compounds, dissection would not be very frequent and would be amply compensated for by the reduction in the size of the mechanical memory and the resulting decrease in access time.

There are, however, two problems which complicate the situation. One is the fact that the semantic content of many constituents differs according to whether they are bound or free forms. The second is that the conventional written form of the majority of the compounds of certain important languages lacks graphic indication of the "seam" between their constituents. Moreover, many compounds permit more than one dissection into constituents identifiable in the mechanical memory. In most cases, however, only one of these is linguistically correct, whilst those in which two dissections are linguistically permissible are extremely rare coincidences. Numerous examples demonstrating these phenomena will be found below.

These complications are such that it seemed at first impossible to create a mechanism which would supply only correct dissections in every case. No wonder Professor Victor A. Oswald, in his paper *Microsemantics* read at the first CONFERENCE ON MECHANICAL TRANSLATION at M.I.T. in June 1952, stated: "We know of no mechanical process by which this could be accomplished, but an intelligent . . . pre-editor could indicate the dissection for any sort of context." The only alternative to the intervention of a human agent seemed to be the inclusion in the mechanical memory of *all* compounds of the source language, an alternative hardly relished by any linguist or engineer. Nor is it humanly possible, as will be seen as soon as we consider the phenomenon of *unpredictable compounding*, customary

¹ This paper is a revised version of my *Studies in Mechanical Translation*, No. 7, September 3, 1952.

in many languages and particularly extensive in German, whose vocabulary is continuously being replenished by this method. Unpredictable compounds can not be coded into the mechanical memory. If no mechanical solution can be found for the problem of the *linguistically correct* determination of the constituents of compounds, then human intervention can not be eliminated from the identification process of MT.

In the following I shall show that there actually is a very simple mechanical solution to the problem presented by unpredictable compounds.

1. *Ascertainable and Extemporized Substantive Compounds.*

For MT purposes we distinguish two kinds of substantive compounds which we abbreviate to "SC":

Ascertainable SC—that is, those which are long established and, therefore, can be located in German dictionaries. Examples are *Kleiderbürste*, *Hochachtung*, *Gehwerk*, *Nachgeschmack*, *Buchstabe*, *Hochzeit*, *Unternehmer*, *Gegenstand*, etc. They could all be entered into the "capital memory." But, as we shall see, a large number of these ascertainable SC can, without sacrificing source-target semantic clarity, be mechanically synthesized out of "memorized" constituents.

Extemporized SC—that is, those which are the result of new free composition, for example *Marsuraniummonopolskandal*. Their potential number is practically infinite. They can, therefore, not be entered into any memory.

2. *The "X-Factor" In German Substantive Compounds.*

A number of SC are characterized by what I call an "X-factor." It is this occurrence of X-factors which presents the main difficulty in the mechanization of the determination of the constituents of SC. X denotes a letter or letter sequence which could be part of the preceding as well as of the following constituent of a SC. See the following examples, some of which have not yet occurred:

The "t" in *Wachtraum* which is either *Wach/traum* (day dream) or *Wacht/raum* (guard room).

The "er" in *Bluterzeugung* which might be either *Blut/erzeugung* (blood production) or

Bluter/zeugung (the begetting of children suffering from haemophilia).

The "in" in *Arbeiterinformationsstelle* which is either *Arbeiter/informationsstelle* (workmen information office) or *Arbeiterin/formationsstelle* (female worker formation office; wrong dissection).

The "ur" in *Literaturkunde* which is either *Literat/urkunde* (man of letters' document; wrong dissection) or *Literatur/kunde* (knowledge or textbook of literature).

The problem becomes more complex when two or more "X-factors" occur in one substantive compound. For example, *Kulturinfiltrierung* which is either *Kult/ur/infiltrierung* (cult earliest infiltration), *Kult/urin/filtrierung* (cult urine filtering; a semantically impossible interpretation) or *Kultur/infiltrierung* (culture infiltration). Such coincidences are comparatively rare, for formal and semantic reasons, and some of the dissections which are possible in terms of forms listed in the dictionary are not likely to prove correct for formal and/or semantic reasons. Thus one would rather say *Allmähliche Durchdringung einer Kultur* or *Beeinflussung einer Kultur* (gradual penetration of a culture) than *Kulturinfiltrierung*. One will find *Arbeiterinnenformationen* (office for the military formations of female laborers) instead of *Arbeiterinformationsstelle*, and *Literatenurkunde* (document of men of letters) instead of *Literaturkunde* because *Arbeiterin* and *Literat*, though they are substantive forms listed in the German dictionary, would not be used as first constituents in these compounds. And *Dichterinbrunst* can only be *Dichter/inbrunst* (poet's fervour), but hardly *Dichterin/brunst* (a poetess' male-animal-like sexual excitement).

Nevertheless, since the only basis for the mechanical determination of the constituents of a SC is the occurrence or non-occurrence of the memory equivalent of an input form in the MT memory, such cases have to be considered in the solution of the problem.

In order to meet these conditions, a solution is suggested here for the mechanical determination of the "seam" or junction between every set of two constituents of a compound. This solution requires a special memory apparatus based on the following considerations:

The primary aim of all translation is access to the meaning of a foreign text. In MT

the primary aim is *quick* access to the meaning. Access time depends largely on storage economy. If in matching every input form the whole store of entries has to be scanned, then access time will play a great role. But if, through the exhaustive utilization of all distinctive graphic features of the different types of source forms (letter sequence, capital initials, occurrence or absence of space, punctuation marks, conventional diacritic marks, etc.) and through the use of a *categorized* storage system, the different types of source forms can be directed to specific sections of the storage system, then the dependence of access time on storage economy decreases in proportion to the increase of categorization.

Consequently, full utilization of all distinctive graphic features of the source text and a categorization on different levels of the storage system are important requirements of this scheme. In planning the contents of the memory I have given precedence to source-target semantic requirements over storage economy wherever possible.

3. The Capital Memory.

One of the facts on which this solution is based is the conventional capitalization in German of the initial letters of all forms occurring immediately after a final punctuation mark, and of the overwhelming majority of German substantive forms and of a number of other forms in all positions (for examples see below). The graphic distinctiveness thus enjoyed by German substantives not preceded by a final punctuation mark makes it easy to direct them immediately to a special memory. But since substantives also occur as first words after a final punctuation mark, certain measures have to be taken to make sure that *all* substantives reach their matching centre via the shortest possible route.

These measures are the dissection of compounds, economy of access time, and considerations of source-target semantics. They make it necessary to divide the German MT memory into a number of sub-memories. One of these sub-memories is the capital memory for the treatment of all substantives.

At this point, it is desirable to consider German words beginning with a capital letter in some detail.

Words With Initial Capital Letter.

The following German forms have initial capitals:

- a) After final punctuation marks (period, question mark, exclamation mark, the colon preceding direct discourse) *all* first words.
- b) In all positions:
 1. All forms of pronouns used in address instead of *du*, and, in letter writing, all pronouns (including *du*) referring to the addressed person.
 2. All adjectives derived from personal names by the suffix *-isch*.
 3. All adjectives, pronouns and ordinal numbers in titles and in historical and geographical names.
 4. All invariable word forms with the suffix *-er*, derived from place names of provinces or federal states.
 5. All substantives with the exception of certain petrified forms and certain forms used in idiomatic expressions.

All words with initial capital letter, other than demonstrative adjectives, pronouns, non-adjectival adverbs, prepositions, conjunctions and interjections are directed to the capital memory. (In a separate paper² I have discussed how they are sorted and how those not directed to the capital memory can, immediately after input, be directed to their specialized memory.)

Special provision has to be made for cases of initial-capital words after final punctuation marks which may belong to more than one form class. A striking example is *Dichter ist der Hahn geworden* which could mean either "The faucet has become *tighter*" or "The cock has become a *poet*." The ambiguity is here due to antiposition which, though not a feature of the normal word order, is fairly frequent in German.

All substantives with initial capitals are treated in the capital memory. Those without initial capitals are, through the combination of this fact with their letter sequence and with the fact that they are preceded by certain types of words, highly distinctive. They can be dealt with by mechanical processes tailored to the different problems they present.

All other initial-capital words directed to the capital memory are first matched there—that

² This subject is treated in some detail in my chapter "The Mechanical Determination of Meaning" in *Machine Translation of Languages*, New York (John Wiley & Sons), 1955.

is, if they occur also as constituents of SC. If, however, no match is found there, they are passed through the remaining memories in a fixed sequence.

4. The Contents of the Capital Memory.

Certain forms are not included in the capital memory, though they may begin with a capital letter. They are:

- a) Extemporized SC.
- b) Ascertainable SC whose target meaning is inferable from the meaning of the target equivalents of their constituents. For example, *Hochland*, composed of *Hoch* (high) and *land* (land). The target meaning of *Hochland* is "highland."
- c) All unproductive constituents which do not occur as free forms; if all ascertainable SC in which they occur are listed in the capital memory. For example, *Ohn* in *Ohnmacht* (fainting fit).

Most capitalized forms are included in the capital memory, as follows:

- a) All non-compound substantives.
- b) Every SC constituent which:
 1. Occurs as a free substantive form. For example, *Zeit* (time) in *Hochzeit* (wedding).
 2. Occurs as a free, though not substantive form, if not all of the ascertainable SC in which it occurs are entered into the capital memory or if it is still productive. An example is, *Hoch-* in *Hochzeit*. *Hochland* will not be "memorized" because its target meaning "highland" is inferable from the meaning of the target equivalents of the constituents, "high" and "land." An example showing the continued productivity of such forms is "grass" in *Grossneptunien* (the world empire on the planet Neptune).
 3. Does not occur as a free form, if not all of the SC in which it occurs are "memorized" or if it is still productive. This rule takes care of all compounding forms such as *Geschichts* (history) in *Geschichtsunterricht* (teaching of history), or *Ur* in *Ureinwohner* meaning "aborigine" (this *Ur-* is not of the same origin as the free substantive form *Ur* denoting the European buffalo) as against *Ohn* in

Ohnmacht.

- c) All ascertainable SC whose target meanings cannot be inferred from the meanings of the target equivalents of their constituents because the juxta-position of those meanings:
 1. does not make sense. For example *Mitgift* (dowry) composed of *mit* (with) and *Gift* (poison).
 2. makes the wrong sense. For example, *Hochzeit*, composed of *hoch* (high) and "Zeit" (time), together "high time," but actually meaning "wedding" or "nuptials." An example showing that the difference can sometimes be very great is *Unternehmer*, composed of *unter*, meaning "under," and *Nehmer*, meaning "taker," the combined form actually means "contractor" or "employer," not "undertaker."
 3. permits multiple interpretation because of the multiple meanings of the target equivalent of at least one of the constituents. For example, *Ein* in *Einverständnis* may mean "in" as in *Eingang* ("ingoing"—that is "entry, entrance") or "one" as in *Einklang* ("unison"). In *Einverständnis* (agreement) it means "one."

5. Source-Target Semantics in the Planning of the Capital Memory.

The rules stated and exemplified in 4 and especially in 4c will prevent a large number of potential source-target ambiguities and nonsensical target results. But there is another potential cause of source-target semantic difficulties. Many SC share a first or second constituent which has only *two* possible meanings, one characteristic of one group of the SC concerned and the other characteristic of the other group. The most satisfactory solution of this problem is as follows:

- a) If the target meanings of all SC involved can be inferred from the meanings of the target equivalents of both their constituents, then we enter the smaller one of the two groups of SC into the memory unless the constituent or constituents concerned are still productive in one of their two meanings. If both groups happen to have an equal number of members, then we choose either one or the other group for "memorization."
- b) If the target meanings of one group cannot

be inferred from the meanings of the target equivalents of *both* their two constituents, then this group is entered.

- c) In all these cases we enter the two constituents of that group of SC which are not "memorized," and the constituent which both groups share is entered into the capital memory with that meaning in the first position it has in that group of SC which are not "memorized," (see e). For example, *Brech-* in *Brech-eisen* (break-iron, i.e., crowbar) and *Brech-stange* (break-stick, i.e., crowbar), etc., means "break," whereas in *Brechdurchfall* (vomit-diarrhoea), *Brechweinstein* (vomit-tartar, tartar emetic), etc., it means "vomit." If the group of SC in which *Brech* means "break" is the smaller one, then we enter all SC of this group and enter the constituent *Brech* in the sense of "vomit" in the first position.
- d) If, as far as such cases are concerned, a constituent also occurs as a free form—that is, if its free form is identical with its compounding form, then there are the following two possibilities:
1. The free form has only that one of the two meanings of its compounding form, which the latter has in the group of SC not entered. The treatment of this case is identical with that of a free form which has the same meaning or meanings as its graphically *identical* compounding form *none* of whose SC are entered, as for example the free form *Arbeiter* and the compounding form *Arbeiter-* or *-Arbeiter*.) In both these cases only the free form needs to be entered. The graphio-mechanical arrangements in the input and matching system and in the capital memory, required to make this possible, will be discussed elsewhere.
 2. The free form has both meanings of its graphically *identical* compounding form or it has more or entirely different meanings. (The question of the common or different origin of the free and the compounding form plays here no role whatsoever.) Here both forms have to be entered. This situation is exemplified by the free substantive form *Ur*, the two graphically identical composing forms Ur^{-1} and Ur^{-2} and the SC containing these

composing forms. The free form *Ur* means "aurochs" (primitive European bison) and occurs as a constituent (Ur^{-1}) only in one SC, *Urochs* (aurochs). The free form of Ur^{-1} belongs to the poetical style and is not commonly used. Wherever else *Ur-* occurs in an SC, it will be first understood to be " Ur^{-2} ." "Extemporizers" will, therefore, avoid forming new SC with Ur^{-1} . They will use the more common synonym *Aurochs* (or, rarer, *Urochs*) instead. Since *Urochs* is thus the only SC in which Ur^{-1} (aurochs) will occur, it will be entered into the capital memory in order to avoid confusion with the highly productive Ur^{-2} . " Ur^{-2} " occurs in a number of ascertainable SC and is still productive. It means "original, earliest, first." The target meanings of one group of the ascertainable SC containing it can not be inferred from the meanings of the target equivalents of their constituents, as, for example, *Urkunde* (document), *Urteil* (judgment). Thus, as far as the problem of Ur^{-2} itself and the group of SC containing it is concerned, the procedure described above, especially in b, will take care of it. But for the solution of the problem presented by the contrast between Ur^{-2} and the free form *Ur* certain graphio-mechanical arrangements are necessary. These can be understood only after a description of the matching procedure has been given and they will be discussed in a separate paper. I should like to say here, however, that these graphio-mechanical arrangements and the solution of the *Ur* vs. Ur^{-2} problem based on them are remarkably simple.

- e) The target meanings of extemporized SC are mostly inferable from the meanings of the target equivalents of their constituents. These constituents are not likely to carry meanings they do not have as free forms or as components of ascertainable SC. But they may carry a meaning occurring only in SC which are "memorized." Therefore, wherever this is the case, the criterion for the choice between the two groups of compounds described in a) can not be their size, but must be the continued productivity of one of the two mean-

ings of the constituents concerned. The group of compounds none of whose constituents is still productive will be coded into the memory. The other group will be excluded and the still productive constituent or constituents will be coded only with the meaning characteristic of this group—which is the meaning in which the constituent or constituents concerned are still productive. Also, if a group of compounds, which has to be “memorized,” because the meanings of their target equivalents can not be inferred from the meanings of the target equivalents of their constituents, has a constituent which is still productive, the constituent has to be “memorized” too.

6. All Possible Types of German Substantive Constituents

We shall now break down German SC, into all possible types of constituents relevant for their determination. Substantive constituents not accompanied by an “X”-factor, I call “trunk” or “T,” the left trunk “LT,” the right trunk “RT.” If the left constituent contains an “X”-factor, it will be denoted by “LTX,” the right constituent containing an “X”-factor by “XRT.” If the left or right constituent occurs in the capital memory, their notation will have the prefix “p” (possible), if they do not occur, it will have the prefix “I” (impossible). Theoretically speaking, this gives us the following types of substantive constituents.

<i>Left</i>	<i>Right</i>
I. PLT	I. PRT
II. ILT	II. IRT
III. P(PLTX)	III. P(XPRT)
IV. P(ILTX)	IV. P(XIRT)
V. I(PLTX)	V. I(XPRT)
VI. I(ILTX)	VI. I(XIRT)

Of these the left and right forms under VI drop out at once because substantive compounds which have the form “I(ILTX) plus I(XIRT)” or in which either the first constituent has the form “I(ILTX)” or the second constituent the form “I(XIRT)” are linguistically impossible *in all languages*. Consider, for example, the following monstrosities concocted from English material: “literatuin” (“literatu-” from “literature” and “-in” from “aspirin, insulin, etc.”) and “reecutive” (“re-” from “resumption,

resource, etc.” and “-ecutive” from “executive”). “I(ILTX) plus I(XIRT)” would then be the English substantive compound “literatuin-reecutive.” If the right constituent is the possible “executive,” then we get the impossible “literatuin-executive”; if the left constituent is the possible “literature,” we would arrive at “literature-reecutive.”

7. All Possible Types of Substantive Compounds With Two Constituents.

Consequently we need consider only the first five alternatives for both the first and the second constituent. This gives us the following 25 theoretical combinations. (For semantic reasons the examples given are partly unlikely to occur.)

		I.
1. PLT	plus PRT	
<i>Senn</i>	<i>idyll</i>	Alpine herdsman’s idyll.
2. PLT	plus IRT	
<i>Senn</i>	<i>dustrie</i>	<i>An impossible compound.</i> The trunk <i>Das-trie</i> from <i>Industrie</i> (industry) does not occur.
3. PLT	plus P(XPRT)	
<i>Senn</i>	<i>inschrift</i>	<i>Senn, inschrift</i> (inscription), <i>Schrift</i> (writing) and also <i>Sennin</i> (Alpine herdsman) occur.
	(Cf. 11a)	
4. PLT	plus P(XIRT)	
<i>Senn</i>	<i>industrie</i>	Alpine herdsman’s industry. The trunk <i>Dustrie</i> does not occur.
	(Cf. 12)	
5. PLT	plus I(XPRT)	
<i>Senn</i>	<i>ingabe</i>	<i>Ingabe</i> does not occur, but <i>Senn, Sennin</i> and <i>Gabe</i> (gift) occur.
	(Cf. 11b)	
		II.
6. ILT	plus PRT	
<i>Insul</i>	<i>halt</i>	<i>An impossible SC.</i> <i>Halt</i> occurs but <i>Insul</i> does not occur.
7. ILT	plus IRT	
<i>Insul</i>	<i>dustrie</i>	<i>An impossible SC.</i> Neither the trunk <i>Dustrie</i> of <i>Industrie</i> nor the trunk <i>Insul</i> of <i>Insulin</i> occurs.
8. ILT	plus P(XPRT)	
<i>Insul</i>	<i>intoleranz</i>	<i>Insul</i> does not occur, but <i>Intoleranz, Toleranz</i> and also <i>Insulin</i> all occur.
	(Cf. 16a)	
9. ILT	plus P(XIRT)	
<i>Insul</i>	<i>industrie</i>	<i>An impossible SC.</i> Both <i>Insulin</i> and <i>Industrie</i> occur, but neither <i>Insul</i> nor <i>Dustrie</i> occur.
	(Cf. 17)	

10. ILT plus I(XPRT)
Insul *ingabe*
(Cf. 16b) Neither *Insul* nor *Ingabe* occur, but *Insulin* and *Gabe* (gift) occur.
- III.
11. P(PLTX) plus PRT
Sennin a) *schrift*
b) *gabe*
(Cf. 35) *Sennin*, *Schrift* (or *Gabe*) all occur. Also *Senn* and *Inschrift* occur, but *Ingabe* does not occur.
12. P(PLTX) plus IRT
Sennin *dustrie*
(Cf. 4) The trunk *Dustrie* does not occur, but both *In-dustrie* and *Senn* occur.
13. P(PLTX) plus P(XPRT)
Sennin *inschrift* Alpine herdsman's inscription. But also *Senn* and *Schrift* occur, though *Sennin* and *Ininschrift* do not occur.
14. P(PLTX) plus P(XIRT)
Sennin *industrie* Alpine herdsman's industry. *Senn*, *Sennin* and *Industrie* all occur, but *Dustrie* and *Inindustrie* do not occur.
15. P(PLTX) plus I(XPRT)
Sennin *ingabe* An impossible SC. *Senn*, *Sennin* and *Gabe* occur, but neither *Ingabe* nor *Sennin* nor *Iningabe* occur.
- IV.
16. P(ILTX) plus PRT
Insulin a) *toleranz*
b) *gabe*
(Cf. 8 & 10) Insulin tolerance or insulin gift. *Intoleranz* occurs, *Ingabe* does not occur; the important fact is, however, that *Insul* does not occur.
17. P(ILTX) plus IRT
Insulin *dustrie*
(Cf. 9) An impossible SC. Both *Insulin* and *Industrie* occur, but neither *In-sul* nor *Dustrie* occur.
18. P(ILTX) plus P(XPRT)
Insulin *information* Insulin information. *Insulin*, *Information* and *Formation* all occur, but *Insul*, *Insulinin* and *In-information* do not occur.
19. P(ILTX) plus P(XIRT)
Insulin *Industrie* Insulin industry. Neither *Insul*, *Dustrie*, *Insulinin* nor *Inindustrie* occur.
20. P(ILTX) plus I(XPRT)
Insulin *ingabe* An impossible SC. *Insulin* and *Gabe* occur, but neither *Insul*, *Ingabe*, nor *Insulinin* occur.
- V.
21. I(PLTX) plus PRT
Steinin *schrift* *Steinin* does not occur, although *Schrift* occurs. But both *Stein* and *In-schrift* occur.
22. I(PLTX) plus IRT
Steinin *sel* Both *Steinin* and *Sel* do not occur, but *Stein* (stone) and *Insel* (island) occur.
23. I(PLTX) plus P(XPRT)
Steinin *inschrift* An impossible SC. *Stein*, *Inschrift* and *Schrift* occur, but neither *Steinin* nor *Ininschrift* occur.
24. I(PLTX) plus P(XIRT)
Steinin *insel* An impossible SC. *Stein* and *Insel* occur, but neither *Steinin* nor *Ininsel* occur.
25. I(PLTX) plus I(XPRT)
Steinin *ingabe* An impossible SC. *Stein* and *Gabe* occur, but neither *Steinin* nor *Iningabe* occur.
- Of these 25 combinations 2, 6, 7, 9, 15, 17, 20, 23, 24 and 25 are linguistically impossible. Of the remaining 15 combinations, 3 and 11a, 4 and 12, 5 and 11b, 8 and 16a, and 10 and 16b represent the same SC; 3 and 11a present, moreover, two possible dissections of the same SC (i.e. *Senn/inschrift*, Alpine herdsman's inscription, and *Sennin/schrift*, Alpine herdsman's writing). Thus only 5, 8, 10, and 12 can be ignored. This leaves us with the following eleven possible types of SC:
- 1,3,4
11 a & b, 13, 14
16 a & b, 18, 19
21 and 22.
- Of these eleven types only two types with an identical graphic form, 3 and 11a, are ambiguous. From the point of view of the matching mechanism these two types are only one type, so that only ten types remain. Thus only in one out of ten possible types will the matching mechanism have to supply a double answer. (But see "Compounds With An X-Factor," section II, below.) In all other cases the answer will be unique. Furthermore, since all the unique answers and the one double answer are obtained in one to four matching steps, the remaining ten types present only four possible matching situations with which the design engineer has to deal. For these I refer to Section 10, below.

8. Matching Procedure for Substantives Which Have A Complete Memory Equivalent And For Substantive Constituents.

As we have seen in 4, only free substantive forms and productive substantive constituents are entered into the capital memory. Substantive constituents which also occur as free, though not substantive, forms are entered only as compounding forms. Thus the "substantivized" adjective *Rot* (*Das Rot der Vorhänge passt nicht zur Farbe der Teppiche* "the red of the curtain does not suit the colour of the carpets"), the compounding forms *Rot* (*Rotstift*, red crayon), *-gelb-* and *"grün"* (*das Rotgelbgrün der bolivianischen Handelsflagge* "the red-yellow-green of the Bolivian merchant flag"), and *Mit-* in the sense of "co-" (*Mitarbeiter*, *Mitbesitzer*, *Mitbürger*, co-worker, co-owner, co-citizen) etc., will be entered, but not the free adjective forms *rot*, *gelb*, *grün*, *hoch*, nor the free preposition form *mit*. These will be entered in their own specialized memories. On the other hand SC like *Mitgift* and *Mittag* would be "memorized."

The capital memory is subdivided into sections characterized by the number of component minimal symbols (space and letter symbols) of entries. Thus entries with five minimal symbols will be in the five-symbol section, entries with four symbols in the four-symbol section, and so forth. Within each section the order is alphabetical. The input mechanism counts the minimal symbols of each form fed into it and directs those forms which have not previously been directed to other memories² at once to the capital memory section indicated by the number of symbols.

Such an arrangement will go far to cut down the access time: substantives are checked only against the capital memory, and within the capital memory only against memory equivalents with the same number of letters. If the memory counterpart of a substantive form does not occur in the section characterized by the number of its symbols, the matching mechanism ignores the last symbol and checks the remainder against the section with the next smaller number of symbols. This process is repeated until the first agreement is found. The sequence of symbols previously ignored is then fed back as a new input and sub-

jected to the same process until the memory equivalents of all substantive components have been located. The constituents established by this process are individually translated in their original sequence.

All substantives not found as complete entries or determined through the matching process described above appear on the target side in their original form.

In the following each completed matching procedure will be called "one matching step."

9. Matching Procedure For Mechanical Determination Of Constituents Of All Substantive Compounds.

I. Left To Right Matching.

P(PLTX)

A. If RT has no memory equivalent, (*Sennin/IRT* P(PLTX) IRT *dustrie, Schülerin/vasion*, cf. 7/12), then the matching mechanism feeds back LT (*Senn, Schüler*, male student) and XRT (*Industrie, Invasion*) and determines the memory code for LT and XRT.

P(ILTX)

B. If RT has a memory equivalent, (*Insulin/PRT* P(ILTX) PRT *toleranz, Insulin/gabe*, cf. 7/16), then the matching mechanism feeds back LT (*Insul*) and,

ILT

1. if LT has no memory equivalent, (*Insul/P(XPRT)* ILT P(XPRT) *intoleranz, Insul/ingabe*, cf. 7/8,10), then the matching mechanism supplies the memory code for LTX (*Insulin*) plus RT (*Toleranz, Gabe*).

PLT

2. If LT has a memory equivalent, (*Stein/P(XPRT)* *inschrift*, cf. 7/21), then the matching mechanism feeds back XRT (*Inschrift*) and,

PLT

a) if XRT has no memory equivalent, (*Senn/I(XPRT)* PLT I(XPRT) *ingabe, Wäscher/inzeichen*, cf. 7/5), then the matching device supplies the memory code for LTX (*Sennin, Wäscherin*, laundress) plus RT (*Gabe, Zeichen*, mark).

PLT

- b) If XRT has a memory equivalent, (*Senn/P(XPRT)* *inschrift*, cf. 7/3 and 11a), then the matching mechanism has to supply two answers: the memory code for LTX plus RT (*Sennin/schrift*) and for LT plus XRT (*Senn/inschrift*).

II. Right-To-Left Matching.

Note: *Left-To-Right* matching presents the simpler engineering problem. *Right-To-Left* matching has the advantage that it tackles first the final constituent which can only be the compounding form of an existing or non-existing (cf. “-nahme” in “Landnahme” land taking) substantive and contains all the grammatical information there is about the SC in which it occurs.

ILT

- A. If LT has no memory equivalent, (*Insul/P(XPRT)* *ILT P(XPRT)* *intoleranz*, *Insul/ingabe*, cf. 7/10), then the matching device feeds back LTX (*Insulin*) and RT (*Toleranz*, *Gabe*) and determines the memory code for LTX and RT.

PLT

- B. If LT has a memory equivalent, (*Senn/P(XIRT)* *PLT P(XIRT)* *industrie*, *Schüler/invasion*, cf. 7/4), then the matching mechanism feeds back RT (*Dustrie*, *Vasion*) and,

P(PLTX)

1. if RT has no memory equivalent, (*Sennin/IRT P(PLTH)* *IRT* *dustrie*, *Schülerin/vasion*, cf. 7/12), then the matching mechanism supplies the memory code for LT (*Schüler*, *Senn*) plus XRT (*Invasion*, *Industrie*).

I(PLTX)

2. If RT has a memory equivalent, (*Steinin/PRT* *schrift*, cf. 7/21), then the matching mechanism feeds back LTX (*Steinin*) and,

- a) if LTX has no memory equivalent, I(PLTX) PRT

(*Steinin/schrift*), then the matching device supplies the memory code for LT (*Stein*) plus XRT (*Inschrift*).

- b) If LTX has a memory equivalent, P(PLTX) PRT

(*Sennin/schrift*, cf. 7/11), then the matching mechanism has to supply *two* answers: the memory code for

LT plus XRT (*Senn/inschrift*) and for LTX plus RT (*Sennin/schrift*).

10. Number of Matching Steps Necessary for Mechanical Dissection of Substantive Compounds with Two Constituents.

The matching mechanism always determines first the longest memory equivalent. We are here concerned with the number of matching steps of only those SC which do not occur in the capital memory. We distinguish the following possibilities:

- No constituent occurs in the memory.
- Only one constituent occurs in the memory.
- Both constituents occur in the memory.

Those with only one or no constituent occurring in the capital memory are at once directed to the output print system and put out in their source form as are all other words not found in the memory.

For SC both of whose constituents occur in the capital memory we distinguish between:

- Compounds without an “X”-factor.
- Compounds with an “X”-factor.

In the following only “left-to-right” matching will be considered.

The examples represent *types* of compounds. They need not actually occur.

Compounds Without An “X”-Factor

For compounds without an “X”-factor (i.e. *Nach/geschmack*, “after-taste,” *Senn/idyll*, “Alpine herdsman’s idyll”; cf. 7/1) we receive a unique answer after the last letter (in right-to-left order) of the second constituent (that is, the *g* of *-geschmack* and the *i* of *-idyll*) has been ignored by the matching mechanisms—that is, *after the first matching step*. The determination of *Nach-* and *Senn-* as largest memory equivalents—that is, as first constituents—determines *-geschmack* and *-idyll* as second constituents.

Compounds With An “X”-Factor

I. Compounds Always Yielding A Unique Answer

A. After The First Matching Step

Compounds yielding a unique answer after the *first* matching step because the form with first trunk plus “X” (*Steinin-* in the following examples) does not exist.

The following facts can be ignored by the machine and the memory designers:

- The second trunk exists:

Steinin-schrift (Cf. 7/21. Solution: *Stein/inschrift*, stone inscription.)

2. The second trunk does not exist:

Steinin-sel (Cf. 7/22. Solution: *Stein/insel*, “stone island.”)

B. After The Second Matching Step

Compounds yielding a unique answer after the *second* matching step because the *second trunk* (*-dustrie*, *-vasion* in the following examples) does not exist.

The following facts can be ignored by the planners:

1. The first constituent has only one “X”-factor:
Sennin-dustrie (Cf. 7/4. Solution: *Senn/industrie*, “Alpine herdsman’s industry.”)
2. The first constituent has two “X”-factors:
Arbeiterin-vasion (Solution: *Arbeiter/invasion*, “workmen’s invasion.”)

C. After The Third Matching Step

Compounds yielding a unique answer after the *third* matching step because the first trunk (*Insul-* in the following examples) does not exist:

1. There is only one “X”-factor between the two trunks. The following facts can be ignored by the planners:
 - a) The second trunk can not have an “X”-factor prefix (*-ingabe* in the following example does not exist):
Insulin-gabe (Cf. 7/16b. Solution: *Insulin/gabe*, “insulin gift.”)
 - b) The second trunk can have an “X”-factor prefix (*-intoleranz* in the following example exists):
Insulin-toleranz (Cf. 7/16a. Solution: *Insulin/toleranz*, “insulin tolerance.”)
2. There are two identical “X”-factors between the two trunks. The following facts can be ignored by the planners:
 - a) The second trunk (*-dustrie* in the following example) does not exist:
Insulin-industrie (Cf. 7/19. Solution: *Insulin/industrie*, “insulin industry.”)
 - b) The second trunk (*-formation* in the following example) exists: *Insulin-information* (Cf. 7/18. Solution: *Insulin/information*.)

D. After The Fourth Matching Step

Compounds yielding a unique answer after the *fourth* matching step because the form with “X”-factor plus second constituent (*-ingabe*, *-inindustrie*, *-ininschrift* in the following examples)

does not exist:

1. There is only one “X”-factor between the two trunks:
Sennin-gabe (Cf. 7/5. Solution: *Sennin/gabe*, “Alpine herdsman’s gift.”)
2. There are two identical “X”-factors between the two trunks. The following facts can be ignored by the planners:
 - a) The trunk of the second constituent (*-dustrie* in the following example) does not exist:
Sennin-industrie (Cf. 7/14. Solution: *Sennin/industrie*, “Alpine herdsman’s industry.”)
 - b) The trunk of the second constituent (*-schrift* in the following example) exists:
Sennin-inschrift (Cf. 7/13. Solution: *Sennin/inschrift*, “Alpine herdsman’s inscription.”)

II. Compounds Yielding A Double Answer After The Fourth Matching Step Unless the “Ur”-Problem Solution Is Incorporated In the Matching Mechanism.

Compounds all of whose trunks (*Literat* and *Welt* in the following example) and forms with *trunk plus “X”-factor* as well as *“X”-factor plus trunk* (*Literatur* and *Urwelt* in the following example) occur in the capital memory, but whose left trunk (*Literat*) does not occur as a left constituent of SC, would, unless the “UR”-problem solution (cf. 5/Db) is applied, yield a *double answer* after the *fourth* matching step.

Such compounds are, for formal and semantic reasons, rare coincidences:

Literatur-welt:

Solution a) *Literatur/welt*, world of literature—correct dissection.

Solution b) *Literat/urwelt* literary man’s primeval world—wrong dissection.

Since *Literat* cannot be a first constituent, the *Ur*-problem solution is applicable and a *unique answer* will be supplied by the matching mechanism after the *third* matching step: the compounding form *Literat-* will not be found in the capital memory.

The case of the following Russian example is similar:

rybo-lovu

Solution a) :*rybo/lovu*, to a fisherman—correct dissection.

Solution b) :*ryb/olovu*, to the tin of the fishes—wrong dissection.

Both trunks *ryb* (genitive plural of *ryba*, fish) and *-lovu* (compounding form meaning “to a catcher”; cf. *ptitse/lovu*, to a fowler, and *kryso/lovu*, to a rat-catcher), and also the composing form *rybo-* (a trunk-plus-“X”-factor form) and the free form *olovu* meaning “to the tin” (an “X”-factor-plus-trunk form) will occur in the capital memory. The connective vowel *-o-* is an “X”-factor. But the trunk *ryb* cannot be a first constituent and the compounding form *ryb-* will, therefore, not be found in the capital memory. Consequently, the matching mechanism will supply a unique and the correct answer after the third matching step.

III. Compounds to Which the “Ur”-Problem Solution Cannot Be Applied and Which, Therefore, Always Yield a Double Answer After the Fourth Matching Step.

For compounds in which two dissections are formally correct and semantically valid, the “Ur”-problem solution is not applicable. These will, therefore, always yield a double answer after the fourth matching step. Such composita are, however, extremely rare coincidences:

1. “*Sennin-schrift*”

Solution a): *Sennin/schrift*, Alpine herdsman’s writing (cf. 7/11a).

Solution b): *Senn/inschrift*, Alpine herdsman’s inscription (cf. 7/3).

2. “*Wacht-raum*”

Solution a): *Wacht/raum*, guard room.

Solution b): *Wach/traum*, waking dream, daydream.

In such cases the MT mechanism will supply two alternative translations.

11. The Mechanical Dissection of Substantive Compounds With More Than Two Constituents.

The solution for the mechanical dissection of SC with two constituents includes the solution for the mechanical dissection of SC with more than two constituents. For the matching mechanism such composita are nothing but SC with two immediate constituents, namely the largest first signal sequence which has a memory equivalent, plus the rest. Once the longest first signal sequence with a memory equivalent is established, the matching mechanism feeds back the rest, and

the procedure is repeated until all constituents are determined.

Let us assume that all non-compounded constituents of *Grieselbärintelligenzexperiment* occur in the capital memory. The first longest signal sequence with a memory equivalent established by the matching device will then be *Griesel-* (grizzly), and *Bärintelligenzexperiment* will be fed back. Note the “X”-factor *-in-* after *Bär*. *Bär* means “bear,” *Bärin* “female bear.” The first longest signal sequence now established will be *Bärin*, and *-telligenzexperiment* will be fed back. Since no portion of this rest can be found in the memory (*-telligenz* does not exist), the matching device will feed back *Bär* (cf. 9/I), locate its memory equivalent and feed back *Intelligenzexperiment*. It will now establish *Intelligenz* as the first longest signal sequence occurring in the capital memory and *Experiment* as the last constituent. Solution:

Griesel/Bär/Intelligenz/Experiment,

Grizzly bear intelligence experiment.

12. Vocabulary Research: Lexical Information Required.

The solution suggested in the preceding pages for the mechanical determination of the constituents of *all* substantive compounds indicates the type of qualitative and quantitative lexical information required for the planning of the capital memory and the matching mechanism. The most important points of this information are:

1. How many and which non-compound substantives, substantive compounds and non-substantive forms belonging to the general language, or only to a specialized language, are eligible for the capital memory?
2. How many and which ascertainable SC can be “synthesized” without any loss in source-target semantic clarity?
3. How many signal number sections will be necessary? What will be the number of source forms in each section?
4. How many and which eligible forms *are unproductive, have been productive, are still productive: are or are not “X”-factor forms; have non-distinctive, distinctive or both types of composing forms; can only occur as left constituents (cf. Lehr-), or only as right constituents (cf. -lehre, -kandidat, -nahme), or as both*

(cf. *Arbeiter*-, *-arbeiter*); which forms cannot, or are not likely to, occur as *constituents* of *proper names* of *source language origin* (cf. *Erziehung*, education, *Verwundung*, wounding, *Tisch*, table, *Sessel*, chair, etc., etc.); which forms only occurring as *right constituents* are *not listed in dictionaries* (cf. *-nahme*)?

5. In how many and which cases do the *free* and the *non-distinctive* compounding forms have the *same*, *different*, or only *two meanings*, one carried only by the *free*, the other only by the compounding form?
6. In how many and which cases does the compounding form have the *same meaning* in *all SC* in which it occurs (cf. *Arbeiter*-, *-arbeiter*); when does it have *two meanings*, one associated with *one*, the *other* with a *second group of SC* in which it occurs?
7. How many and which *SC* *permit double dissection*? To how many and which ones can the "*Ur*"-*problem solution* be *applied*, i.e.:
 - a) How many and which "X"-factor forms have a "possible" trunk or an "impossible" trunk?
 - b) How many and which "X"-factor forms occur with the same "X"-factor?
 - c) How many and which "X"-factors occur?

I may add here that some "X"-factors are, for morphological reasons, of frequent occurrence (for example *-er*-, *-in*- and *-ur*-); others, for formal and semantic reasons, are rare (for example the *-t*- in *Wachtraum*).

"X"-factors can be easily located in the vocabulary by determining whether, after one or more final or initial letters of a productive or potential substantive constituent are dropped, the remaining letter sequence represents another productive or potential substantive constituent. Examples are the finals and initials in *Wacht* (guard), *Wach*- (waking), *Traum* (dream), *Raum* (room), in relation to *Wachtraum*; *Traum* (dream), *Trau*- (wedding), *Mahnung* (exhortation), *Ahnung* (foreboding), in relation to *Traumahnung* (dream foreboding); *Lehrer* (teacher), *Lehr*- (teaching), *Erzeugnis* (produce), *Zeugnis* (certificate), in relation to *Lehrerzeugnis* (teacher's certificate); *Bärin* (female bear), *Bär* (male bear), *Instinct* (instinct)—containing an "impossible" trunk *-stinkt*—in relation to *Bärinstinkt*; *Kultur* (culture), *Kult* (cult), *Urwelt* (primeval world), *Welt* (world), in relation to *Kulturwelt* (civilized

world).

8. Since all German words after a final punctuation mark have a initial capital letter, vocabulary research will also have to determine all ascertainable substantives whose graphic form—apart from the initial capital letter—is identical with that of a form belonging to another form class.
9. Another important category which should be established in the course of this vocabulary research is all two-initial-letter combinations possible in the source language and the size of the membership in each combination group. To go beyond the second initial letter would not be practical because three-letter words are frequent. The membership of each signal-number section of the capital memory could then be further subdivided into groups of source forms with the same two-initial-letter combinations. The matching mechanism would then compare each source form only with those memory equivalents in the signal-number section concerned which have the same two-initial-letter sequence. This procedure would further reduce access time to a degree where it would be negligible from the MT point of view.

13. Conclusion

The mechanical identification—demonstrated here for the German language—of all compounds which are not included in the mechanical memory and lack graphic indication of the boundaries between their constituents is, of course, applicable to other languages. Only minor modifications in the mechanical design and in the programming will be necessary to take care of differences in the graphic distinctiveness of form classes, such as the absence of the capitalization of substantives, other than proper names, in non-initial positions. Other minor adjustments in this scheme will make it possible to eliminate from the mechanical memory most free and bound forms of dual nationality which has been treated separately.

The importance of the mechanization of this part of the identification process of MT lies in the fact that it solves the problem of unpredictable compounds and makes possible a substantial reduction in the size of the mechanical memory with a resultant decrease in access time. The compound effect of these results in the lowering of the cost of MT is obvious.