[*Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University, 1957]

Word Decomposition for Machine Translation

R. H. Richens and M. A. K. Halliday
Cambridge Language Research Unit, Cambridge, England
Presented by R. A. Crossland, University of Durham, England

All feasible systems of machine translation are based on a unit smaller, in a great many cases, than the word. This unit, which provides the source-language entries in a mechanical dictionary, is conveniently termed a "chunk" so as to avoid confusion with other linguistic categories. There are, however, a number of ways in which words may be decomposed for machine-translation purposes and the following remarks deal with some of the principles that are involved.

Linguistic structure

By applying the well-known method of testing for linguistic commutability, it is possible to establish classes of chunks each characterized by its internal commutation relations. All the chunks considered below satisfy commutation tests; but since in many cases this technique gives a number of possible word decompositions, it is necessary to consider which possibilities should be adopted. In addition, some systems at least of machine translation require principles of word decomposition based on quite other criteria than those used in structural linguistic analysis.

Invariant words

There is no particular difficulty with invariant words such as English prepositions. Here the word is a chunk. Yet it may be that such a chunk forms a semantically irreducible compound with some other chunk or chunk class. Thus *up against* and *have (has, had, having) up* need to be treated as couplets whose meaning cannot be inferred from the normal range of meaning of the component chunks. It is simplest, however, to treat *up* in each case as a chunk whose meaning can only be elucidated by comparisons with neighbouring chunks. This comparison may have to await syntactic analysis as in *hurry him up* where *hurry* and *up* form an irreducible but disjunct semantic compound.

Affixation

A single affix likewise presents few difficulties. Both linguists and machine translators will decompose *dogs* into *dog-s* and *unkind* into *un-kind.* The notion of affixation, however, contains implications

that must be investigated further, in particular the implied distinction between stem and affix. It is certainly the case that this distinction is commonly made on semantic grounds; the stem is regarded perhaps as an argument, operated on by its affix, or at least as having some sort of semantic priority to it. The distinction could be maintained purely on formal linguistic criteria with reference to the extension of the commutation classes, but it is probable that most methods of machine translation utilize the stem-affix distinction and it does not greatly matter for machine-translation purposes whether the distinction is purely formal or formal-semantic.

The question then arises whether a segment which can be either a stem or an affix, such as *or* and *-or,* is to be regarded as one chunk or two. In the case of two-chunk words, the position of the space bounding the word can be used to provide a basis for distinction; but this becomes more difficult with three-chunk words such as *possess-or-s.* It could be maintained on structural grounds that *-or* and *-or-* are one chunk and *or* another; alternatively *-or* and *-or-* could be regarded as different chunks. The position is even more complex with regard to segments such as *en, en-* and *-en* which can function as stem, prefix or suffix. Moreover, both *en-* and *-en* can be infixed as in *dis-en-thrall* and *moist-en-s.*

To avoid merely verbal dispute, it is convenient to regard all segments composed of the same letters in the same order as one and the same chunk. However, there is no reason why the stem and affix distinction or distinction into different classes of affixes should not be applied within the chunk if the system of machine translation being used requires it. Thus the *-en-* in disenthrall can be regarded as (1) a prefix, (2) an infixed prefix or (3) an infix, each being a subcategory of the affix category of the *en* chunk. It is possible also to classify an affix by numbering its distances either from the beginning or end of the word or from the stem. Thus the *-en-* of *disenthrall* is in position 2 from the beginning, 2 from the end, and 1 before the stem. Any of these methods can be justified on structural analytic grounds; the one adopted will depend on the machine translation method used, and different methods might well require different ways of classifying affixes.

Multiple affixation involves further problems. Commutability considerations would permit *disenthrall* to be divided either as *dis-en-thrall, disen-thrall* or *dis-enthrall.* It is not even necessary for

all parts of a chunk to be contiguous. Something is to be gained by decomposing the German *abgeschrieben* into *ab-, ge- -en, schrieb;* and similarly in the Semitic verb. However, it is probably simpler to regard disjunct affix couplets as consisting of two different chunks which form an irreducible but disjunct semantic compound as discussed above. It is doubtful whether any consideration based on formal-linguistic analysis alone can be adduced for any particular system of decomposition when multiple affixation occurs. The decision has in fact to be made on the various other criteria mentioned below.

Mutable stems

The simplest instance is exemplified by initial mutation in Welsh where mutation occurs without affixation, e.g. *pen, ben, mhen, phen.* There are two possibilities here, either to regard these variants as four different chunks or to divide into *p-en, b-en, mh-en, ph-en.* Since most Welsh words are affected by initial mutations, the increase in the size of the mechanical dictionary if all variants were entered as separate chunks would be very serious. On the other hand, if *en* is regarded as a unitary chunk, it is necessary to be able to distinguish the above series from *c-en, g-en, ngh-en, ch-en;* moreover *g-en* is not only a variant in the *c-en* series, but may be a root form of the series *g-en, en;* moreover *en* is itself a root form. It is clear that the removal of the mutable initial from the stem is only possible if the root form is recoverable at a subsequent stage in the mechanical-translation procedure. This is easily done, as pointed out by Richens (1956) by treating 9 Welsh initial mutation series as flexional classes, adding this information to the mechanical dictionary and then comparing stem and initial letters for flexional class. The mode of decomposition adopted must, therefore, depend on the flexional system set up.

Mutation plus affixation

Combined mutation and affixation is more complicated. An example is provided by *half, halves.* Obvious possibilities for decomposition are as follows:

half	*halves*
half	*halv-es*
hal-f	*hal-ves*
hal-f	*hal-ve-s*
hal-f	*hal-v-es*

Here again, economy of mechanical index space will tell against the first two possibilities. The choice between the others will depend on the subsequent choice of a flexional system. The last form, *hal-f, hal-v-es,* has the advantage that it can be integrated easily with the verbal set.

hal-v-e, hal-v-es, hal-v-ing, hal-v-ed

Reduplication

Reduplication plus affixation is very similar to the preceding. Thus the series

lop lops lopping lopped

admits of several treatments, in particular

lop	*lop-s*	*lopp-ing*	*lopp-ed*
lop	*lop-s*	*lop-pin g*	*lop-ped*
lop	*lop-s*	*lop-p-ing*	*lop-p-ed*
lo-p	*lo-ps*	*lo-pping*	*lo-pped*
lo-p	*lo-p-s*	*lo-pp-ing*	*lo-pp-ed*

As before, the first possibility is likely to be discarded since it involves two stem entries in the dictionary. Choice between the other forms will depend on the flexional system used subsequently. The third possibility is probably the best for a number of systems of machine translation.

Semantic requirements

In the foregoing, word decomposition has been treated almost entirely as an exercise in linguistic analysis. For machine translation, however, it is frequently necessary to sacrifice a commutatively possible division if there is no semantic parallelism. Thus while *in-exact* and *in-excusable* are usable decompositions, *in-famous* or *in-fam-ous* is not, since *infamous* does not mean *not famous.* Prepositional prefixes to verbs in Teutonic and Romance languages offer many other instances. It is simplest to treat *infamous* as a single chunk.

Flexional system

It has been noted already that the system of decomposition depends on the flexional system. The following table illustrates two of the many possibilities of treating a constellation of words containing the segment *cop.*

Word	System I Division	Flexional class	System II Division	Flexional class
cop	*cop*	b	*cop*	p
cops	*cop-s*	b	*cop-s*	p
copping	*copp-ing*	b	*cop-p-ing*	p
copped	*copp-ed*	b	*cop-p-ed*	p
copper	*copper*	b	*copper*	b
coppers	*copper-s*	b	*copper-s*	b
cope	*cope*	b	*cop-e*	e
copes	*cope-s*	b	*cop-es*	e
coping	*cop-ing*	b	*cop-ing*	e
coped	*cop-ed*	b	*cop-ed*	e
copious	*copious*	a	*copious*	a
copy	*copy*	a	*cop-y*	y
copies	*copies*	a	*cop-ies*	*y*
Total stems	7		3	
Total affixes	3		8	
Total flexional classes		2		5

a = invariant stems; b = stems with simple affixation; c = stems affixing *-e;* p = stems infixing *-p;* y = stems affixing *-y.*

The main advantage of system I lies in its economy in affixes and the small number of flexional systems. A serious drawback is the large number of stems, more than double that in System II. The second system is economic in stems but at the cost of 8 affixes and 5 flexional systems. However, since the number of affixes in any system is limited while the number of stems is roughly proportional to vocabulary size, system II is probably far more economic of mechanical-dictionary space for a large vocabulary. If dictionary size is not important, system I may be simpler to manipulate.

Chunk identification

In addition to the limitations on decomposition set by the conjugation system, further limitations may be set by the mechanical matching technique. Most methods of machine translation envisage comparison of the words of the source passage with the chunks in a mechanical dictionary, and since a large number of words contain several chunks, techniques have to be devised to identify these

chunks correctly. There are doubtless many ways of doing this, but it is probable that most methods impose restrictions on certain semantically and structurally permissible decompositions where otherwise misidentifications will occur.

An example will make this clear. One of the simplest techniques of matching (cf. Richens and Booth, 1955) is to match each word, beginning at the front end, against a mechanical dictionary of chunks arranged in alphabetic order but with the longer words preceding the shorter. Then, when a match is made, i.e. a chunk in the mechanical dictionary corresponds exactly with an initial segment in the word, the remaining segments, if any, are rematched. This method is simple and works in a great many cases. thus *disloyalty* will be decomposed as follows:

disloyalty dis loyalty
loyalty loyal ty
ty ty

However *discontent* by this method would be liable to yield:

discontent disc ontent
ontent on tent
tent tent

This can be prevented by applying the standard solution of so many machine-translation problems, namely by putting the cause of trouble in the mechanical dictionary. Thus, *discontent,* though semantically separable into *dis-content,* is treated as the unitary chunk *discontent.*

It is possible to avoid trouble with this particular word by using a different matching technique, but it is likely that any comparatively simple technique will result in misdivision in some cases. This is of no consequence if it is clearly recognized that any awkward word is to be treated as a unitary chunk.

Translation field

Machine translation schedules may be classified into 9 categories, according to whether they go from or to the particular, comparative or universal (cf. Halliday, 1957); that is whether there are one, several or all source languages catered for, and one, several or all target languages. Thus, a scheme which applies only to English-Italian is an example of one-one translation; a general programme to render any language into a Romance language would be an all-several translation.

Word decomposition for machine translation may depend to some extent on the translation field. In translation between related languages, it may be possible to utilize nonsemantic parallelisms due to common origin or borrowing. Thus the infix *-iz-* in English has a number of quite different meanings, e.g.

sympath-iz-e	to manifest sympathy
pulver-iz-e	to bring to powder
mechan-iz-e	to do by machine

Parallels to these words exist in French, Italian and Rumanian:

sympath-iz-e	*sympath-is-er*	*simpat-izz-are*	*simpat-iz-a*
pulver-iz-e	*pulvér-is-er*	*polver-izz-are*	*pulver-iz-a*
mechan-iz-e	*mécan-is-er*	*meccan-izz-are*	*mecan-iz-a*

The parallel uses in these four languages are such that it is feasible to some extent to decompose as above and to translate English *-iz-* by Italian *-izz-* etc.; even though these infixes vary widely in meaning within each language. On the other hand, in translating any of the above into Japanese, this type of decomposition would be less appropriate as the divergent significance of the infix in the examples quoted requires a different rendering in each case.

The above illustration represents a relatively clear instance of a chunk which may be translatable comparatively but not universally. It can be maintained as axiomatic that no affix is universally, one-all translatable; that is, in linguistic terms, there can be no universal identification for translation purposes of any grammatical category. For this reason especially, there are obvious advantages in one form of decomposition for any one source language, whatever the target language or the translation field. If there is no possibility of the translation of a given segment as a chunk it is probably best handled at a later stage in the machine-translation programme. For example, there is a category of “plural” in part of the Chinese noun-system, but it cannot be arrived at by simple translation from the English plural, i.e. translation of the English -s chunk; this does not necessarily mean that in an English-Chinese translation programme the *-s* plural should not be handled as a chunk, but that it would not be represented directly in the Chinese, and the Chinese category of plural would have to be introduced by other means.

Conclusion

The object of the preceding note is to show that the range of possible word decompositions can be established by commutation tests as carried out by normal structural linguistic analysis. The actual decomposition appropriate to any particular situation can only be decided by additional criteria. Of these, semantic requirements, flexional system, chunk identification technique and translation field are the ones that have been considered; but other criteria are not excluded.