ANNETTE STACHOWITZ

BEYOND THE FEASIBILITY STUDY: LEXICOGRAPHIC PROGRESS*

1. FEASIBILITY STUDY

The feasibility study on fully automatic high quality translation was held in 1971 under the auspices of the Linguistics Research Center of The University of Texas at Austin with support from RAFB. The participants were twenty experts in the areas of linguistics and computer software, representing a wide range of opinions on the feasibility of MT. Among the recommendations made in the report on this study was one regarding the need of " lexical research to determine the syntactic and semantic patterns of linguistic entities ". The result of such research would be a dictionary containing, in addition to lexical information, syntactic and semantic features and restrictors necessary for quality translation. In this paper, I would like to describe some of the work which is currently being done at the Linguistics Research Center toward such a dictionary.

2. WORK AT THE CENTER

2.1. *Approach.*

2.1.1. *Purpose.*

The purpose of the dictionary is to provide the information which the grammar of a language can refer to in its rules to

a) establish the wellformedness of a sentence,

b) avoid forced reading, and

---

* This paper is an updated version of a paper presented at the Tenth Annual Meeting of the Association for Computational Linguistics in Chapel Hill, North Carolina, 1972.

c) recognize the semantic reading of each word in a particular context (for correct choice of one translation equivalent from a number of possibilities or for synonymy substitution in paraphrasing).

This information should be as comprehensive as possible; on the other hand, it must still have manageable proportions. Our aim is to find the happy medium; our classification system constitutes a first modest step towards it.

2.1.2. *Types of Information.*

In general, the features we include in our lexicon reflect surface phenomena: there are several reasons for this approach.

To begin with, the input text for any translation consists of surface strings; whatever information we need is present in these surface strings and must be recognized.

Secondly, it will be simpler to recognize regularities and establish transformational relations as well as semantic classes once we have comparatively comprehensive lists of lexical items with all or at least most of their surface features. Any hypothesis can then be checked against this set of data. When we describe surface phenomena, we do not overlook distinctions like those between *easy to please* and *eager to please* or *expect somebody to do something* and *advise somebody to do something*, etc. These pairs of lexical items share only *some* surface environments, not all of them, a fact which we can recognize when we compare their descriptors, by inspection or mechanically.

Further, a description of surface structure gives us the possibility of producing translations which are structurally similar to the input string if these are permissible and of similar stylistic value in the target language.

In addition to surface features, our description includes some of the transformational features which are well established. For example, we mark verbs which cannot be passivized, English verbs which do not form the progressive, and adjectives which modify the verbal aspects of a noun rather than its full scope of meaning.

Finally, we include in our lexical description some indications of functional relationships between surface and deep structures. These are based on a contrastive approach; they are included only for those sets of translation equivalents for which these relations are not identical. For example, the English verb *fail* takes a subject and a complement

both of which are basically the same in the surface and the deep structure. Its German translation, the verb *mißglücken*, however, takes a subject which appears as a dative object in the surface, while the surface subject represents the deep object complement. This characteristic of the German verb is explicitly marked in our lexicon.

2.2. *Methods.*

The method used at the Center for the compilation of its MT dictionaries consists of three steps:

1) inclusion of information provided in existing dictionaries;
2) coding of carefully selected semantic and syntactic features;
3) interpretation of the information in the resulting lists and revision.

2.2.1. *Information which is provided by existing dictionaries.*

Among the available dictionaries we found most useful are the *German-English Dictionary* by WILDHAGEN and HÉRAUCOURT and *The Advanced Learner's Dictionary of Current English* by HORNBY, GATENBY and WAKEFIELD. Both of these dictionaries indicate the different readings of a lexical item separately, according to syntactic and/or semantic criteria. Therefore, our initial lists already contained a small amount of syntactic and semosyntactic information, such as human vs. non-human subjects or objects required by verbs.

We began with the German and English verb lists primarily to find out the general types of selection restrictions of these verbs and to use this information as the basis of our classification of nouns and adjectives. In this way, we compiled the following lists:

a) A German verb list with English translation equivalents. This list contains approximately 17,500 entries. Since the different readings of each key-word constitute separate entries, the list represents approximately 12,000 different verb stems.

b) An English verb list, consisting of 6,500 verb stem entries.

c) A German noun list with English translation equivalents, gender, and inflectional information. This list contains appr. 73,000 entries; at present, appr. 2,000 entries contain semosyntactic features.

d) A bilingual adjective list of appr. 27,000 entries, each consisting of an English keyword, its German translation equivalents, and subject area or stylistic descriptors. The addition of features to these entries is still in progress; approximately 12,000 items have been assigned features so far.

e) An English-German adverb list which is still in the process of compilation. At present, it covers adverbs beginning with the letters A through R, totalling some 3,500 entries. Each entry consists of an English adverb, its German translation equivalents, and syntactic and semo-syntactic features as described under 2.4.5. in this paper.

f) In addition to these lists of one-word entries, we have compiled bilingual lists of phrasal verbs. The lists are separated according to internal syntactic constituents:

$V + PRPH$ — phrases, such as *take into account = in Betracht ziehen* (4,500 entries)
$V + NP$ — phrases, such as *raise a question = eine Frage stellen* (5,700 entries)
$V + A$ — phrases, such as *aehnlich sehen = look like* (900 entries)
$V + V$ — phrases, such as *zu verstehen geben = indicate* (700 entries).

2.2.2. *Updating of lists.*

The second and current stage in the lexicographic work is the addition of syntactic and semo-syntactic features not explicitly given in the existing dictionaries but identified as essential by linguistic research of the past two decades. The information is being coded according to a general classification system, which I will touch on briefly later and for which we rely heavily on the intuition of native speakers.

2.2.3. *Interpretation of lists.*

The third stage of our lexicographic work is the interpretation of the information contained in our lists. For this purpose, we produce concordances in which all lexical entries having a particular feature in common are listed together. Even more useful will be the sub-lists which our linguists can request – lists of all lexical entries with specific combinations of features, for example. English verbs which can take both

that-clauses and infinitive complements, or German reflexive verbs whose English translation is intransitive, etc. Inspection of such special sub-lists will, hopefully, reveal regularities and semantic sub-classes which we can use to replace combinations of syntactic or semantic features and thus streamline the dictionary.

Similarly, those features which can be predicted from the presence or absence of others will be deleted from the dictionary and introduced by lexical redundancy rules. An example would be a redundancy rule stating that all verbs which take *for-to* object complements require human subjects.

2.3. *The Classification System.*

The classification system according to which lexicographic work at the Center is being done is essentially the same for German and English. Some syntactic structures, however, apply only to one language. Basically, the system comprises three types of features: the properties of the classified element itself; the properties of the environment in which it may occur wellformedly; and information pertaining to the subject area to which a lexical element or one of its semantic readings might be restricted.

a) The features of the element itself are its syntactic category, its possible discontinuous elements, such as separable verbal prefixes in German and adpreps in English, and its semantic sub-classification.

b) The selection restriction of the classified element indicate the syntactic and semantic restrictions on obligatory and optional complements, mostly of verbs, adjectives, and prepositions, but also of some nouns and adverbs.

The features of the lexical element and also its selection restrictions are indicated as values of subscripts, the latter being generalizations of the features. For example, *TC*(*AB*) stands for " this verb requires an *NP* complement which is abstract ". *Or*, as in " requires a human *or* abstract complement " is indicated by a comma: *TC*(*HU, AB*). The comma also indicates ambiguity; for example, the English noun *operator* has the features *TY*(*HU, AB*), which is to be read as " this noun is ambiguous; it is abstract in one reading, human in another ". Optionality of complements is expressed by the value *LA* (lambda) with a comma, as in *FC*(*A, LA*), which reads " this verb (or adjective) takes an optional complement noun phrase in the accusative ".

c) The additional restrictions consist at present mainly of tags indicating the subject area for those words which are restricted to or have special usage in such areas as zoology, law, music, etc.

2.4. *Specific Features.*

At this point I would like to outline briefly what the specific syntactic and semo-syntactic features are which are, at present, being assigned to lexical entries.

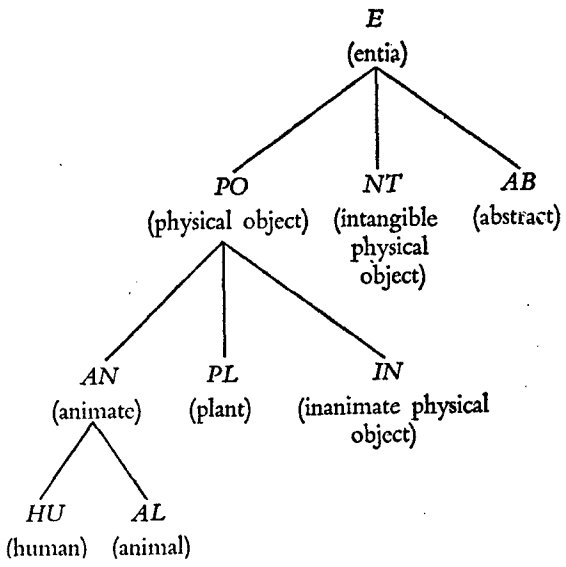
2.4.1. *Nouns.*

The list of nouns contains mostly the features of the nouns themselves and only few selection restrictions. They are subclassified according to the requirements of verbs, adjectives, and prepositions, but not according to logical distinctions or a subclassification of the universe. In other words, we are not interested in real distinctions but in the distinctions made in the language. The two are not always the same, as the classification of the English noun *novel* might show. A novel is normally considered an abstract, yet we have to classify it, in addition, as inanimate physical object because one speaks of a novel meaning the book or stack of papers on which it is printed. Thus we say: *Hand me that novel. Don't sit on my novel!* etc.

The basic semantic noun classes we have established are abstract, human, animal, plant, and inanimate physical object. To these, a number of additional markers can be added as relevant. These are: title (for items like *Mr.*, *Dr.*, etc., which can be followed by proper names); name (for frequently occurring proper names); body part (because of verbs like *amputate* and because the German definite article preceding nouns denoting body parts is replaced by a possessive pronoun in English); machine (because many verbs which normally require a human subject can also take a machine as a subject); mass (because these may be used without an article in the singular); collective (because such nouns can be used with verbs which otherwise require plural subjects or subjects, as e.g. *disperse: the crowd dispersed, they dispersed,* but not **the man dispersed*; in English such nouns are frequently referred to by plural pronouns: *The government is in agreement;* ***they*** *are expected to release* ***their*** *decision soon.*); quantity (for nouns like *glass, half, per-*

cent, *dozen*, etc., which can occur in the frame a —— (of) *NP*, where they act as the syntactic but not the semantic head of the noun phrase construction); time (a sub-class of abstract; these nouns can take a *when*-relative clause); count (abstracts which are preceded by a definite article, such as *idea*, *event*, as opposed to those abstracts which may be used in the singular without an article, e.g. *mathematics*, *love*, etc.; the latter type is unmarked); unit (abstracts like *mile*, *year*, etc., which fit into the frame $NP = QUANT + ——$, the whole noun phrase being a quantitative adverb, as in ***five miles*** *long*; this allows us to analyze only these noun phrases as adverbs, instead of considering every *NP* a possible adverbial, which would cause a large number of incorrect readings).

Some nouns also have selection restrictions, mostly prepositional objects or clauses. Examples are *dependence* (*on*), *fact* (*that*). The required syntactic form of such complements is indicated as *that*-clause, marked or unmarked infinitive, gerund, *for-to* phrase, interrogative adverb plus clause, specific prepositions, and possible combinations of these. For prepositional objects, the semantic type of the noun within the prepositional phrase in also indicated as one of the semantic noun classes mentioned earlier. To indicate semantic selection restrictions, the basic semantic noun classes are combined in tree form as shown:



Each node of the tree may be used as a selection restriction in order to save enumeration of all noun types which are permitted for a particular complement. Sub-classes of adverbs are also used as semantic selection restrictions.

Since we have no derivational analysis as yet and therefore cannot mechanically recognize the underlying verbs of deverbative nouns, we have one additional noun feature, indicating optional directional adverbs which may function as complements of some noun; example: *the trip to Peking, the flight to the moon.* This feature is needed for correct analysis and prevention of forced readings in strings like *They described the* ***trip to the South Pole.*** vs. *They* ***transported*** *these items* ***to the South Pole.,*** where the adverb modifies the noun in the first, the verb in the second sentence.

2.4.2. *Verbs.*

Verbs and adjectives are assigned mostly selection restrictions. Only two features in our verb lists describe the verb itself. The first of these is *PX* (for " prefix "), in which slot we spell out the specific separable prefix or adprep with which a verb may occur. All other features refer to the combination verb stem + prefix or adprep. The second verb feature which is not a selection restriction is " type ", which only indicates, redundantly, whether the verb is transitive, intransitive, or reflexive; also indicated are the small classes of transitive verbs which may not be passivized and of English verbs which do not form the progressive.

Verb selection restrictions are syntactic form and semantic type of subject and objects required; multiple objective are indicated by a " + " and given in the order direct object – indirect object. All possible types of complements are indicated except, of course, the free ones, such as the benefactive dative, adverbial infinitive phrases, etc. (For a list of syntactic complement markers, cf. p. 329; verbs and adjectives may also have features indicating *NP* complements, with case indication in German). For example, the English verb *please* is coded as follows:

V
ON(C)
FS(N;CL)
TS(E;TH,MI,ICL,FT,GR)
FC(NP,LA;NP)
TC(HU,LA;HU)

The features indicate that this verb may occur with a noun phrase or a clausal subject; if the subject is a noun phrase, it may be of any semantic type and the verb takes an optional noun phrase object of the semantic type " human "; if its subject is a clause, it may be a *that*-clause, a marked infinitive, an interrogative adverb followed by a clause, a *for-to* construction, or a gerund; in any of these cases, the verb takes an obligatory human object.

In addition to these selection restrictions, verbs which require an adverbial complement have a feature indicating the specific type of adverb, as e.g. the feature " requires an adverb of direction " for the English verb *put*. Verbs which may optionally occur with an adverb of direction are given a special marker. It may be of interest that this class of verb is not identical with the semantic class of motion verbs because it includes such verbs as *extend*, *look*, etc.

Finally, we have the two features " interpretation of surface subject " and " interpretation of surface object ", which indicate such surface – deep structure relations as in the pair of translation equivalents *fail* – *mißglücken*, which I described earlier.

2.4.3. *Phrasal Verbs.*

The problem of phrasal verbs, of course, cannot be solved in surface analysis, with a surface dictionary, since they can occur discontinuously and in various permuted word orders, and since their meanings, translations and features cannot be derived from their individual components. Our surface dictionary only tags the one-word verbs and the nouns which may form part of such a phrasal lexical item. The full verb phrases are entered as lexical units in our verb phrase lists and given the same types of features as normal one-word verbs.

Some of these phrasal items have internal slots, usually possessive pronouns. These are indicated by different descriptors within the phrasal entry, denoting reflexive and non-reflexive possessive pronoun slots. Typical examples with such variables are

(etw.) ***zu jds.*** *Kenntnis bringen = notify, inform* ***sb.*** *(of sth.)*

with a possessive pronoun slot, and

seinen *Anfang nehmen = begin*

with a reflexive possessive pronoun slot.

These lexical entries are not meant for surface analysis but are applied to analyze re-ordered versions of surface strings, the so-called " standard strings ", in which the components of phrasal elements are contiguous and occur in a pre-determined order. In this analysis, such phrases are assigned the features and translations which pertain to the whole phrasal unit.

2.4.4. *Adjectives.*

As indicated earlier, the only semantic classes of adjectives which we mark are tough movement adjectives and those which modify only the verbal aspect of the noun, as described by ZENO VENDLER in his *Transformational Grammar of English Adjectives.* A typical example of adjective feature may be given in the English adjective *eager*:

A	=	*EAGER*
ON(V)		
FM(NP)		
TM(AN)		
FC(CL,LA)		
TC(MI,LA)		

The features given are to be read as: " This adjective must take a subject in the form of an *NP* which belongs to the semantic subclass ' animate '. It takes an optional complement in the form of a marked infinitive ", as in *He is eager. He is eager to please.* Since a marked infinitive subject is not permitted by this description, we prevent analysis and generation of **To please him is eager.* This automatically puts *eager* into a different class from that of *easy*, whose description includes marked infinitive subjects.

2.4.5. *Adverbs, Prepositions and Conjunctions.*

Our classification system for adverbs, prepositions and conjunctions is, as yet, very tentative. It includes classes like place and time, place being subdivided into static, direction to, and direction from, time into punctual, duration, and frequency for some items, present, past and future for other items, and, where relevant, into prior to, simultaneous, and posterior to. Some other classes are modal, divided into manner,

comparison, and restrictive; degree, cause, purpose, instrumental. These classes are used for adverbs, prepositions, and conjunctions. Adverbs have, in addition, a feature indicating whether they may modify verbs, adverbs, numbers, noun phrases, adjectives, or full sentences; if they modify only declarative sentences or only questions or only negated sentences, this fact is indicated. We also classify adverbs like *rather* or *clearly* as being restricted to co-occurrence with certain types of subjects or verbs, respectively, *rather* requiring an animate subject, *clearly* (as verb modifier) a verb of mental activity. In general, we try to indicate all restrictions we can formulate.

Prepositions have, in addition to semantic type, a feature indicating the semantic type of noun with which they are used. This information may serve to select one of several readings of the preposition, as in the case of the German preposition *nach*, which indicates direction if it occurs with a noun of the type " location name ", as in *nach Europa* (*to, towards Europe*), time if it occurs with any type of noun, as in *nach dieser Konferenz* (*after this conference*) and reference if it occurs with a human noun, as in *nach Dr. von Braun* (*according to Dr. von Braun*).

3. CONCLUSION: PROJECTED WORK

I hardly need to point out that the lexicographic work I have described is only a start. Much more work is necessary, guided by the experience we are gathering along the way and that of others who are working on various problems in lexicography and linguistics in general.

Our most immediate task is extending our classification system, especially for the semantic classification of verbs and adjectives. Study of various sub-lists of our initially classified lexical data should be very helpful in this work.

Also of great importance is the establishment of restrictive glosses where syntactic and semantic features and our present very general " subject area restrictors " are not sufficient. Thus, we will have to include information which is rarely given explicitly in conventional dictionaries. Wildhagen, for example, sometimes gives in parenthesis for verb entries sample subjects or objects which indicate to the human reader what general kind of subject or object must be used. This type of information is sometimes the only way to distinguish the different readings, and therefore, translations or synonyms, of a lexical item.

Consider, for example, the German verb *einhalten*. If its object is cloth or something made of cloth, its translations into English is *gather*; if its object is a movement, a gesture, etc., its translation is *check*; with an object like law, contract, date, diet, anything which signifies a rule of some kind, its translation is *observe*; with path, way, course, etc. it is translated as *follow*. The information which lets us choose the correct one from all these readings must be incorporated in the selection restrictions of the verb and in the feature system of the nouns.

Currently, our word lists include derivational forms and compounds. We are planning a study of these items to determine to which extent their syntactic and semo-syntactic features can be derived from those of the underlying roots and affixes. The development of productive derivational and compounding rule systems would, of course, result in a considerable reduction in the size of the dictionary.

The lists resulting from our work will be appended to our future progress reports. Two intermediate lists have already appeared in our report on *Research in German-English Machine Translation on Syntactic Level* of August, 1970 and in *Normalization of Natural Language for Information Retrieval*, 1971. A comprehensive description of our lexicographic work is contained in our two most recent reports, *Development of German-English Machine Translation System*, of July 1973 and September 1973.

We hope that our lexicographic work will not only result in the construction of an MT dictionary but will also provide information which may prove interesting and useful to linguists in general.

REFERENCES

A. S. HORNBY, E. V. GATENBY, H. WAKEFIELD, *The Advanced Learner's Dictionary of Current English*, Second Edition, London, 1963.

W. P. LEHMANN, R. A. STACHOWITZ, *Research in German-English Machine Translation on Syntactic Level*, Volume II, RADC-TR-69-368, Linguistics Research Center, The University of Texas at Austin, 1970.

W. P. LEHMANN, R. A., STACHOWITZ, *Normalization of Natural Language for Information Retrieval*, Final Technical Report, Linguistics Research Center, The University of Texas at Austin, 1972.

Z. VENDLER, *Transformational Grammar of English Adjectives*, in « Transformations and Discourse Analysis Papers » No. 52, University of Pennsylvania, Transformations and Discourse Analysis Project, Philadelphia, 1963.

K. WILDHAGEN, W. HÉRAUCOURT, *English-German, German-English Dictionary*, Volume II, German-English, Wiesbaden, 1953.