Meta
Journal des traducteurs
Translators' Journal

META

Terminological Records and Lexicon Entries. A Contrastive Analysis

Ingo Hohnhold et Thomas Schneider

Volume 36, numéro 1, mars 1991

La terminologie dans le monde : orientations et recherches

URI : https://id.erudit.org/iderudit/003076ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

Découvrir la revue

Citer cet article

Hohnhold, I. & Schneider, T. (1991). Terminological Records and Lexicon Entries. A Contrastive Analysis. *Meta*, *36*(1), 161–173.

Résumé de l'article

Les entrées des systèmes de traduction automatique doivent contenir un ensemble de traits et de valeurs linguistiques plus élaboré que les entrées des lexiques traditionnels. Après une brève description de ces derniers, on explique l'élaboration des entrées des lexiques dans les systèmes de traduction automatique, leurs fonctions et leur fonctionnement, en plus de rendre compte de certains de leurs traits caractéristiques et des méthodes d'analyse grammaticale et de transfert des structures qu'ils utilisent. Enfin, l'importance de la notion de concept dans le transfert des stuctures grammaticales est soulignée et on termine avec des réflexions sur la place des terminologues et du travail lexicographique dans le monde de la traduction automatique.

TERMINOLOGICAL RECORDS AND LEXICON ENTRIES. A CONTRASTIVE ANALYSIS.

INGO HOHNHOLD AND THOMAS SCHNEIDER
Siemens AG, Munich, Germany

RÉSUMÉ:

Les entrées des systèmes de traduction automatique doivent contenir un ensemble de traits et de valeurs linguistiques plus élaboré que les entrées des lexiques traditionnels. Après une brève description de ces derniers, on explique l'élaboration des entrées des lexiques dans les systèmes de traduction automatique, leurs fonctions et leur fonctionnement, en plus de rendre compte de certains de leurs traits caractéristiques et des méthodes d'analyse grammaticale et de transfert des structures qu'ils utilisent. Enfin, l'importance de la notion de concept dans le transfert des stuctures grammaticales est soulignée et on termine avec des réflexions sur la place des terminologues et du travail lexicographique dans le monde de la traduction automatique.

1. REVISITING A LONG-DECIDED QUESTION?

In the past, lexicon entries had usually been defined as a reduced version of the underlying terminological records. The transition of a terminological record to a lexical entry simply implied the deletion of certain items of information. This concept applied — and applies — to both conventional specialized dictionaries and dictionaries and glossaries derived from terminology data bases. (Section 3 will discuss the validity of this assumption).

Entries in the dictionaries of machine translation systems, by contrast, require a fundamentally different approach. They do leave out partial information contained in terminological records, but they contain sets of linguistic features and values which go beyond the limits of terminological records. This type of lexicon entry exhibits in this regard an increase of informational content over a terminological record.

Thus lexicon entries in conventional dictionaries intended for human translators and lexicon entries in machine translation systems differ greatly. Conventional entries don't have to contain formal linguistic data since a human translator will usually be able to use the entries properly without such information. Missing elements can generally be inferred by intuitive rule application or by analogy. Today's state of the art in machine translation however requires a system to contain the linguistic information explicitly since "human" aspects such as drawing analogies cannot be modelled and implemented with sufficient reliability.

The points mentioned suggest a new and more detailed investigation of lexical entries as such. Significant connections within terminological lexicological and lexicographic work may easily be overlooked if the conclusion drawn is limited to the simple statement that machine translation has produced a new type of entry.

2. TERMINOLOGICAL RECORDS

In this paper, terminological records are used simply for the sake of comparison but they are not described in detail as this has been done in a different publication[1]. Readers not familiar with definition and content of terminological records might review the main aspects for a better understanding of the points made here.

The differentiated view of lexical entries proposed here does not affect the terminological record, neither in its role as the fundamental unit of any terminology work nor in its substance. But it is conceivable to link terminological records and lexicon entries in machine translation systems in the future with the intent of reducing redundancy and duplication of effort. This is further explained in section 5.

3. LEXICON ENTRIES IN CONVENTIONAL TRANSLATION

Lexicon entries in conventional translation are assumed to include all word or concept entries in dictionaries and other collection (*e.g.* terminology data bases) which are utilized as a direct means in the translation of specialized texts; for reasons discussed later, entries in dictionaries for machine translation are to be treated separately.

3.1 SOURCE

All these entries share the common trait in principle and usually in practice they are subsets of originally much larger clusters of items of information. To put it differently, they are a condensed result of previous lexicographic or terminological research and description which is presented in a lexicographic format. This explains the previously mentioned view of a lexicon entry as a reduced set of information. In the following analysis, we will trace how such lexical entries originated, the reasons for their formal condensation, their content and functions, and we will point out the difference to terminological records.

Processes in the Creation of an Entry

If a collection of subject matter information under the heading of an entry, as for example a terminological record in a data base, is to be turned into a useful lexical entry, two processes can occur: either the simple deletion of partial information or abstraction.

Deletion means that specific items of information contained in the source are left out without substitution by any other items. In this case, it is possible that the sum of items of information within a specific category is reduced, as perhaps only one of the original three listed synonyms is taken over into the lexical entry. Or perhaps a complete category is ignored, as *e.g.* all examples of the use of the term in context or the source of the term may be deleted.

The process of abstraction intends to draw conclusions from available partial information and to insert them into the lexical entry in the shape of generalized statements. Often, concrete linguistic descriptors are turned into subject matter information. To give two examples: The sources listed in a series of context examples suggest that the term is typical for a specific type of text; this information appears in the lexical entry. Or the definitions in the two languages of a bilingual terminological record show that there is no complete equivalence of the two concepts but that nevertheless the two terms can or must be accepted as corresponding terms; in the lexical entry, a note would point out the limited equivalence.

In addition to the partial information selected from the source or generated by abstraction, there may be other items included in the lexical entry which were not derived from the source, for example grammatical descriptors or phonetic transcriptions.

Of course, the described processes are closely interwoven. They do not usually occur in isolation. A useful lexical entry will not exist until the transition of the material and possibly its enlargement has produced a unit of independent status. The creation of

valid lexical entries is never simply a reduction of existing material; selection and condensation are complex processes.

Reasons for the Condensation of the Material

It is quite obvious that dictionaries, especially technical dictionaries, can be marketed successfully only if the intended target group, *i.e.* translators, can afford them. Therefore dictionaries cannot be too voluminous and, as a consequence, expensive. This goal can be achieved in one of two ways. Either the dictionary is designed to contain a large number of entries, each with limited information content, or it is structured so as to have few but extensively described entries. In practice, almost exclusively the first possibility is chosen, since a reduced set of entries would limit the market drastically, and publication would be a gamble for author and publisher.

There are legitimate reasons why dictionaries with a wide coverage of larger subject fields are so popular. It is not just that translators have carried over the habit of using one or two standard dictionaries from the translation of general texts. But the types of technical texts which are to be translated usually contain terms from the whole subject field and often also from neighboring fields even if the text rates as highly specialized. Considering the limited time translators are permitted for their task, it is imperative that a large number of the terms in question be found at the first consultation. In other words, one dictionary has to cover most of the terms the translator is looking for.

The generalizing assumption that a translator doesn't need more than just the term in question and its equivalent in the target language as a lexicon entry is untenable. To be sure, it does happen that a translator, especially an experienced specialized translator, does not need more than the "equation" in specific cases but as a rule this does not hold true. For instance, terms for complex concepts may be used differently in the target language, but this may not be visible at the word level. Or a term within a certain subject field may denote a concept which differs slightly from one in a neighboring field expressed by the same term, and this might have an effect on the interpretation of the context and its translation.

One point however is valid: a lexicon entry in a dictionary compiled for translators should be as short and precise as possible and give the most important information since translators usually don't have the time to plow through pages of accompanying information. But here, too, it would be unwise to generalize. Highly complex matter may require a detailed description.

3.2 CONTENT

One distinctive feature in dictionaries is the basis of the entries: an entry may be word-based or concept-based. This distinction has a more significant effect on the content of the entries than any other aspect. In a word-based entry, the word in question is listed in its different meanings or in its most important meanings, each with its target-language equivalents. Synonyms are usually not listed. By contrast, a concept-based entry refers to a single — invisible — concept which appears in the shape of its denotation. Words of identical appearance but with a different referent (polysemes) are not a necessary element of the entry and are usually not included; rather, they form separate entries. Instead, synonyms are listed. Word-based entries are typically found in general-language dictionaries while specialized dictionaries require concept-based entries.

In the following we will focus on the concept-based entry in a technical dictionary which has been condensed from a terminological record. This is the type of entry with which a technical translator deals most of the time, and which is closely related to his terminology work. Its relevant content can be best described by reviewing the informational units of a terminological record as cited in Hohnhold (1988) and stating if and to which extent they can be used for lexical entries.[1]

In both cases the first and foremost position is occupied by the terminological unit. This may be a single or multi-word entry, a contiguous phrase or an idiom of discrete elements; sometimes it may occur in a common inflected form instead of the canonical form. It constitutes the head of the entry and is used as the key in look-up. Multi-lingual entries have a corresponding number of target language equivalents. Phrases can usually be found listed alphabetically under one or more of the keywords occuring in the phrase.

Closely related to the terminological unit are synonyms such as orthographic variants, abbreviations or international scientific labels. This information is contained in only a small number of entries but where it exists it is of prime importance to a user. Antonyms can sometimes be useful since they tend to clarify a concept by means of contrast. Lengthy and abstract definitions are likely to be unwieldy, and examples for the use of a term in context may not be listed at all. The same applies to an indication of the source.

Extremely important is the indication of the subject field. This information can only be dispensed with if the whole dictionary deals with just a single narrow subject field like *e.g.* chromatography or automobile transmissions. Denotation of the concept is necessary only if the dictionary contains a systematic part in addition to the alphabetical part. References to the semantic environment of the concept may be helpful in individual cases, especially for beginners.

Information about limited range of applicability within the subject field is mainly relevant within the framework of a company. If the limitation is general in nature it should be listed. References to text type or level of speech and to geographical distribution are especially important with synonyms. If a term is considered rare or outdated this fact ought to be listed, too.

In cases where there is no complete equivalence between corresponding terms in different languages an indication of the degree of equivalence is indispensable. It would be quite irresponsible to mislead a casual user who, without warning, might take the correspondence of the terms listed for granted.

The date of acquisition or first input of a record is irrelevant. Instead, the dictionary should clearly indicate how valid and up to date it is[2].

3.3 FUNCTION

The lexical entries considered so far have one essential function. They have to inform the user rapidly and reliably about meaning, use and validity of denotations and their target-language equivalents. It can be assumed that the user with their creative ability to select and combine information and to draw analogies will actually enlarge the content of an entry beyond the explicitly stated items.

3.4 DISTINCTIVE FEATURES

If we define the distinction between a lexical entry and a terminological record it has to be viewed with a reservation which is of little consequence in practice but which should be mentioned for the sake of conceptual correctness: A lexical entry in a specialized dictionary basically remains a terminological record, more precisely, a terminological record which has been treated according to lexicographic principles. The concepts of "terminological record" and "lexical entry" reside on different levels. Firstly, one is content-based, the other is form-based. Secondly, terminological records refer to specialized concepts while lexical entries may refer to general concepts as well as specialized concepts. In spite of this, the distinction between terminological record on one hand and lexical entry on the other is common and is retained in this paper.

A terminological record has two main functions, namely to be both a body for the collection of additional up-to-date items of information and a source of information. The

lexical entry has only the second function. At the same time, requirements concerning the use of the information have increased. Terminological records are collections of material "in progress", lexical entries are finished and streamlined products.

Terminological records document information. Lexical entries offer the results of such documentation as an aid in a productive environment. Terminological records contain units of information in their entirety and physically, while lexical entries are confined to conclusions drawn from this information. Terminological records are properly generated in the process of comparative text research and contain small samples of original text passages, *e.g.* in the form of a term in its wider context, wich can sometimes be inserted directly into a translation. In short, they originate in the text and offer the translator to some degree direct ways to the target text. Lexical entries don't usually have this quality; especially extensive phrases tend to get lost completely. With lexical entries, research more often needs to be done elsewhere to decide on a specific translation problem[3].

Finally the most fundamental difference: The terminological record is the classic organizational unit of terminology work, of research. It is created in the phase of collecting the material; in all subsequent stages of terminology work it is retained in its structure and filled with additions. In terminological records the whole process of approximation and verification of equivalence can be observed. Multi-lingual terminological records may be set up tentatively until perhaps it turns out that the terms in question are not or not sufficiently equivalent across the language boundaries, and that there is no basis to postulate a single conceptual unit. Terminological records mark work in progress; the process of verification must be completed here. Only then, and no sooner, can it offer specific partial information and starting points for additional new information to the lexicographer generating the lexical entries. For the life cycle of the dictionary, they are final products. Of course, it makes sense to keep terminological records even after they have been utilized in dictionaries. Further foreign-language equivalents may be added at a later date, or changing concepts might require a revision, *e.g.* if an international standard supercedes a national standard.

4. LEXICAL ENTRIES IN MACHINE TRANSLATION SYSTEMS

As machine translation systems have not yet achived widespread acceptance many translators may not be familiar with the function and structure of lexical entries for such systems. However, the rapidly growing number of installations of MT systems suggests that in the very near future the majority of translation departments and translation bureaus will make use of MT systems in addition to their conventional "human" translation. Building appropriate lexicons for such applications will prove a novel experience for translation departments, regardless of whether the department has conducted organized terminology work in the past or not.

Lexical entries for purposes of machine translation constitute a distinct new class of lexical entries. In the following we will attempt to give a general description of their properties. Of course, not all MT systems are created equal, and some of the more detailed observations will not necessarily be applicable to older systems. Still, most of the conclusions should be relevant for practical application of machine translation in general even though the description of lexical features will be based on the METAL system[4].

First of all we must take into account that a single MT lexical entry is much more of a constituent part of the relevant lexicon than an entry intended for conventional translation. The lexicon, or more precisely, the individual lexicon modules are an integral part of the overall system. A human translator may utilize any number of different lexicons, or choose not to consult any if he/she does not have any questions regarding the

lexicon. By contrast, the machine is limited to its own stored lexicon, and each and every word occuring in a text will pose a question which must be answered by its lexicon.

To be sure, sophisticated systems will generate interpretations of compound words based on constituent parts, guess et word class and morphological behavior. Nevertheless, this information is usually derived from available information contained in the lexicon.

4.1 GENERATION

In contrast to conventional lexical entries, the origin of MT system entries can hardly be attributed to an available pool of information. They are usually created individually; their different purpose requires contents which are alien to the terminology work translators are accustomed to. The intended completeness of a system lexicon for example may be achived by analyzing a large number of representative texts for the subject area in question and running comparison checks before incrementally updating the entries.

4.2 FUNCTIONS AND FUNCTIONAL ASPECTS

The different objectives inherent in creating entries in a system lexicon are the direct result of the change of the "user". The one to work with the entries is no longer the human translator with all his quirks, shortcomings and/or genius but a computer, a "machine" (which may also overheat under extended stress but for different reasons and with different consequences for the translation result). This translation result needs to be achived automatically, and a fair share of the outcome is directly attributable to the information contained in the individual lexical entries.

In section 3.3 we had emphasized that a lexicon entry must inform the user rapidly and reliably. An entry in a system lexicon shares these requirements to its functionality. The demand for speed of course is easily met by the computer. The aspect of reliability however needs to be taken a step further. As the user is not a human being with a vast store of world knowledge to decipher fuzziness and resolve conflicts in the supplied information, the entry must be unambiguous. The machine can hardly be expected to reach a viable decision about an utterance to be interpreted if the axioms, *i.e.* the grammatical and semantic information contained in its memory, cannot be relied upon.

While it is true that with today's technology it is possible to infer some information from the context of a given word, such as guessing the word class of a word to be a noun if it occurs between a determiner and a verb, this certainly does not hold true for all cases. Especially questions concerning the degree of equivalence between terms and the adequacy of a transfer cannot be answered automatically.

4.3 SOME ASPECTS OF MT GRAMMARS AND LEXICONS

Legitimate machine translation systems are not simply automated dictionaries. Before the automatic translation of a given sentence can be attempted the syntactic structure and the function of the individual elements have to be analyzed. This is usually done on the basis of some variation of a phrase structure grammar. Such a grammar contains rules which take as their input a sequence of constituents defined as to word class and produce as output a higher level structure. A sample rule might describe a sentence (S) as consisting of a noun phrase (NP) and a verb phrase (VP):

S — NP VP

Other rules will describe how a noun phrase may be legally (=grammatically correct) formed, *e.g.* from a sequence of determiner (DET), adjective (ADJ) and noun (N):

NP — DET ADJ N

Thus, a phrase like "The old car" could be interpreted as a functional unit, *e.g.* as the subject of a sentence "The old car refuses to hurry."

Word Classes and Constraints

The allocation of nouns, adjectives and verbs to their respective word classes is fairly straightforward. "Car" or "house" are easily recognized as members of the set of nouns. Other word classes pose greater problems. If one were to classify "very" as an adverb, it would satisfy the grammaticality criteria of a rule that permits an adverb to modify an adjective, as in "The soup is very hot." The word "very", however, could not replace the true adverb in a phrase like "He glanced benevolently at his secretary." Or how does one deal with infinitives or adjectives used as nouns as in German "Das Verändern der Datei..." (Updating of the file) or "Das Spannende an der Geschichte..." (the exciting thing about the whole story).

It is quite difficult to decide what constitutes a determiner and what doesn't. A phrase structure rule NP — DET N would handle German phrases like "der Mann" or "ein Mann". The grammatical structure "der eine Mann" (the one man) could be dealt with in a rule of the format NP — DET N. But such a rule, without modification, would also permit an ungrammatical structure like "der der Mann".

In other words, the assignment of grammatical classes to words, especially to function words, is not necessarily "natural" but is motivated by the analysis strategies of the grammar. So it seems doubtful if there can be an application-independent lexicon system providing all necessary features for multiple uses in different environments and linguistic formalisms.

Of course, all the sample rules given above would be quite inadequate for any analysis. A rule like NP — DET ADJ N would build a noun phrase out of any sequence of determiner, adjective and noun regardless of whether they actually belong together. For a rule to be restricted to the generation of only grammatical structures, it would need to be augmented by constraints on the individual constituents and their interaction. The rule would have to contain a test section to decide if the elements agree in regard to case, number, gender etc. The information necessary to perform such a test of course has to be carried in the individual lexical entries. Some of this information may be purely morphological such as the inflectional class of an adjective or the irregular plural forms of a noun.

Other necessary items refer to syntactic behavior. It would be very difficult to analyze a sentence without having access to information about the transitivity type of a verb, or about legal frames. A German analysis will have to know that the verb "helfen" takes a dative object rather than an accusative (in contrast to other transitive verbs). An English entry for the adjective "below" must specify that this adjective follows the noun, and the entry for "rightful" would exclude the predicative use of the adjective. The British collective "police" requires a plural verb form (in contrast to French "police"). The grammar rule guiding the generation of the correct output will only fire if the appropriate feature and value pair is coded in the relevant entry. In other words, the entry has to assume functions which in conventional translation work are handled by the human translator (who is unlikely to be aware of the many processes involved in analysis and translation).

Grammar rules often need semantic information about lexical items. Syntactic ambiguities can sometimes only be resolved if the semantic types of the nouns involved are known. In the German sentence "Das Gras frisst die Kuh" (the cow eats the grass) there is a syntactic ambiguity as to which of the two nouns is the subject. The knowledge that the verb is likely to have an animate subject, and that "cow" is animate while "grass" is not, is the prerequisite for a correct analysis. So the linguistic information required to

parse a natural language utterance goes way beyond the morphological level which may sometimes be included in standard bilingual dictionaries for "human" translation.

A legitimate MT system also carries in the system lexicon detailed grammatical information at the syntactic and semantic level which is used by the analysis (or by the generation grammar). The information is usually stored in the shape of feature/value pairs even if at the surface it does not appear as such.

It would go beyond the scope of this paper to describe in detail all the features that need to be coded for a machine translation system. Suffice it to say that such a system requires explicit linguistic information where a human translator applies his knowledge intuitively. Since the grammar does not have access to a "world knowledge" it must rely on the coded information and its validity. Translation quality hinges on the quality of the stored information — much more so than any conventional process.

Another salient feature of a system lexicon concerns the type of vocabulary it holds. Of course it needs to contain all technical terminology occurring in the text to be translated, not just the difficult or the very specific ones. In addition, however, it must contain all the common and the general words occuring in the text, including function words like prepositions and definite articles — quite a novel aspect in comparison with conventional translation work. Carrying a large portion of general vocabulary in the system lexicon increases the problems associated with polysemy which to a lesser degree already exist in specialized terminologies. It is estimated that approximately 40 % of the English vocabulary may be ambiguous as to word class, such as "back" which may be a verb, a noun, an adjective or a verb particle. Very rarely do cases of polysemy reflect across language boundaries; so it is usually not possible to simply pass on the existing ambiguity to the target language. Thus a machine translation system must resolve these cases in the analysis of the source text so that both function and meaning of a given string are clear before the actual translation process.

4.4 TRANSFER LEXICON

The linguistic criteria described above are used for the analysis or the generation of a language. Translation is a different step. From the viewpoint of a machine translation system, translation means to transfer the grammatical structures of one language to equivalent structures in another language and to substitute lexemes of the source language with equivalent lexemes of the target language. Looking at this process with the eyes of a literary translator or a theoretical investigator of the problem of preservation of style and meaning, it may appear mechanical and inadequate. We should keep in mind, however, that machine translation is not intended for literary texts or stylistic filigree but for the rapid translation of texts whose sole purpose in life is to convey information explicitly and rapidly to a user of this information content.

Beyond that, a modern MT system does not rely on one-to-one equivalences. In the METAL system, for example, the modular structure separating analysis, transfer and generation is also reflected in the system lexicon.

The grammatical information necessary for the grammar rules to operate on is stored only once, in the monolingual lexicons. This type of information is required for the analysis or the generation of well-formed utterances within a language but it has little to do with the actual translation process. The grammatical information stored in the lexical entries is used to steer the transfer, to be sure, but it is autonomous and may or may not be utilized. While the monolingual entries are largely word-based, transfer entries reflect more of a concept-orientation.

The conditions for syntactic transfer are defined in the grammar. Transformation rules turn a German embedded clause into an English relative clause or produce the

appropriate sequence of prepositional phrases in a target language. These grammar rules are usually not open to an end user since in a modern system they are not *ad hoc* local condition rules. They are powerful restructuring rules at various levels, and spontaneous or ignorant tampering with them might lead to unintended devastating results.

Grammatical Criteria in Transfer

As the transfer module in the grammar effects the structural transformation, the transfer lexicon effects the mapping at the lexeme level. The selection of lexical equivalents for a target language may be based on two criteria, the linguistic context of the word in question or the subject field of the text.

One of the grammatical criteria for the selection of the proper transfer is the fact that certain ambiguous nouns lose that ambiguity when used in either singular or plural, or when they occur as the subject of a semantically marked verb. While American "pot" may refer to either a cooking utensil or to an illegal but widespread herbal consumer good, the plural form does not exhibit the same ambiguity. Plural "pots" are always legal.

In German or French, the indication or grammatical gender may be used to define the appropriate transfer. "Le tour" (tour) is no synonym for "la tour" (tower), and the neuter "Gehalt" (salary) and the masculine "Gehalt" (content) do differ realistically.

In a sentence like "The bank told me that my account is overdrawn", the position of "bank" as the subject of a verb usually requiring an animate or at least *social collective* agent can be used to exclude an interpretation of "bank" as *e.g.* a shoreline. Of course, a disambiguation would be much more difficult in cases where "bank" occurs in a different position, as in "All morning he watched the bank".

Verbs are notorious in calling for a multitude of different transfers depending on context. Criteria might be the presence of a specific subject, object, prepositional phrase or particle or another instantiation of a specific verb frame.

The German verb "zerlegen" can be translated into English "analyze", if the direct object is a lexeme denoting sentence, or into "dissect" if the object has the semantic features animate or human, or into "disassemble" if the object is concrete. Transfers for verb phrases tend to be more idiosyncratic than those for noun phrases so modern MT systems provide transfer mechanisms that can utilize any feature or set of features. In the case of the German verb "bestehen" the transfer is dependent not just on the presence of a prepositional phrase but on the presence of a specific preposition to produce the English equivalents "consist (of)", "insist (on)" or "exist". It may get even more detailed as with the verb "warten" which not only requires a prepositional phrase with the preposition "auf" but an accusative in the prepositional phrase as well to produce a "wait (for)".

English "eat" in turn has to be coded in such a way that the German transfer can be chosen on the basis of the semantic features of the subject: a human subject usually requires "essen", a non-human animate subject "fressen".

Similar aspects apply to the other word classes. As long as there are grammatical criteria to decide on the appropriate transfer and as long as these features are coded in the system lexicon the number of translation problems in the *mechanical* process is limited.

Transformations

Of course, translation does not affect just individual words, and in many cases phrases need to be restructured grammatically to be acceptable. It had been mentioned before that an end-user does not usually have access to the grammar rules. However, in some systems such as METAL the lexicon interface provides access to some macros which can have far-reaching structural effects.

It is for example possible to turn a subject into an indirect object and the direct object into the subject of the target phrase, as in English "I like the book" to German "Das Buch gefällt mir".

Prepositional phrases may be mapped onto other prepositional phrases, as in "The politician reaches for his manuscript" to "Der Politiker greift nach dem Manuskript", but they may also be turned into direct objects, as in "Der Senator antwortet auf die Frage" to "The senator answers the question".

Sometimes, elements of the source structure need to be deleted in transfer to a certain target language. The prepositional phrase in "to take into account" would have to be excluded from a transfer to German since the target verb "berücksichtigen" already incorporates the concept expressed by the phrase. By analogy, new elements have sometimes to be inserted in a target language expression, as in English "launch" to German "auf den Markt bringen". Even if all the major structural changes are handled by the general grammar, these lexicon-driven transformations give the translator a powerful tool to influence the outcome of a translation. The prerequisite of course is a good understanding of the interaction between grammar and lexicon.

Modular Lexicon Structure

In many cases grammatical criteria are not sufficient to reach the appropriate target equivalent since a given term may occur either without context or with a context which does not permit a disambiguation at the linguistic level. For such cases, the system lexicon structure is of importance.

It is generally assumed that the most frequent 5,000 words in any one of the European languages will cover close to 90 % of any general text. If we were to increase the number of lexicon entries to include the next 100,000 words, coverage would not have significantly improved. Instead, more can be gained from instituing a structure of subject-specific modules.

In the METAL system, there is a hierarchy of lexicon modules. The highest level is a set of function words like conjunctions and prepositions. This is followed by a module of general vocabulary and, below that, a module of common technical vocabulary, that is, of terms which are not specific to any narrow subject field.

From the next layer on down, all users can structure their own hierarchy of lexicons. In theory as well as in practice there cannot be a single classification system for all subject areas and applications. So an end-user of a machine translation system must have the possibility to define modules geared to the requirements of his own texts. For example, someone working in Civil Engineering might need to introduce lexicon subsets for machine tooling or welding, whereas someone in Data Processing might instead require modules for data transmission, software, hardware etc. In METAL, it is also possible to define transfers for a specific project or product, for a certain customer or for a target country (to account for the difference between British and American English, for example).

Before a translation run is started, the most specific modules appropriate to the subject area of the text are marked. If after the grammatical analysis there are several candidates left for the translation of a word, lexicon lookup proceeeds from this most specific module. Only if no entry is found there, the search continues in the less specific modules. Thus it is assured that a specialized text in a certain subject area will get the appropriate transfers and not some general language translation. The German word "Mutter" in an engineering text would be rendered correctly as "nut" and not as the general language "mother".

5. CONCEPT VS. TRANSFER

Lexicon entries in machine translation systems do not carry explicitly the same information as a terminological record. Items like antonyms, definition, context, usage and source are not listed, and synonyms usually constitute a separate entry. This negative

list is easily explained by the fact that the purpose of a system lexicon entry is not to validate terminology but to utilize the result of terminology work which has taken place already. The terminological record is a collection of material with the character of "work in progress" while the system lexicon entry is a highly formalized final product.

It appears from the description above that today's machine translation systems do not operate on the basis of concepts in the same way that terminology data bases do.

Usually, as long as the grammatical descriptors, morphology, syntactic behavior and semantic features are identical, there will be a single monolingual entry for a word denoting different concepts. In this respect the lexicon entry differs greatly from a terminological record. Disambiguation is attempted from grammatical context, but usually only as far as the translation result is affected by the distinction.

A similar contrast exists in the structure of lexicon modules. For the sake of expediency and to avoid redundancy, lexicon modules are not necessarily identical with sets of subject field terminologies. In a terminology data base a record usually contains information on the subject fields in which the concept occurs and in which the term is applicable. Such a record may list several subject fields, or the data base may even carry several almost identical records, the difference being simply the subject field notation. Thus, the term "buffer" might be listed under the subject fields of data processing as well as telecommunications. The advantage of multiple listings is that it becomes possible to generate a subject-specific lexicon without losing entries which may be carried under a different subject field heading.

Machine translation systems are intended to be used on long texts within limited domains. As their purpose is to speed up the translation process great care is taken in system design to alleviate all task for the user that are not adsolutely necessary. Introducing redundancy would be counter-productive. A user's hierarchy of lexicon modules will contain only those modules which actually are applicable to the texts to be translated. All other subject areas are ignored. The system lexicon will contain only the vocabulary which actually does occur in these texts. To that end, before a translation run is started, the text is checked against the system lexicon. Missing words are added to the lexicon before the translation so that incrementally the lexicon will grow — but not in the sense of conceptually motivated systematic terminology work.

Whenever possible, a term is listed in the highest permissible module. Let us assume that a user has a lexicon structure with the following hierarchy:

level 1: Function Words
level 2: General Vocabulary
level 3: Common Technical vocabulary
level 4: Data Processing
level 5: Software / Hardware / Data Transmission / Specialized Systems
level 6: (subset of software:) AI system / Graphics / Data Bases / etc.

The term "address space" could conceivably be listed in each of the modules on level 6, or in the modules on level 5. However, as the purpose of a machine translation lexicon is not to reflect ontological accuracy but to produce pragmatic results, the term "address space" would probably best be listed on level 4. As the translation of the term does not differ from one subfield to another in the area of data processing it is carried in the most general module. Thus a single entry is sufficient, instead of listing separate entries for each of the subfields. Should at any time in the future a new subfield require a different translation an additional entry can be created for that application. But until such requirement arises the coding effort (as well as the storage space) can be saved.

New Approches

The principles of lexicon organization outlined above refer to natural language processing systems primarily intended for translation. "Deep" analysis is not attempted beyond the steps necessary to produce an adequate output in the target language. Even here, some qualifications have to be made. There is no linguistic theory available today which would be able to describe even a single language completely and unambiguously. So even when the application of machine translation is limited to factual texts of little aesthetic merit there are areas of linguistic ambiguity which cannot be solved by syntactic analysis and a limited semantic apparatus. So the "adequacy" of the output is a rather debatable concept.

For the future, however, the "shallow" analysis strategies in MT systems will probably be supplemented with more extensive semantic components. A Fair amount of research and development is expended on extending the applications to intelligent information retrieval. Conventional systems so far have usually operated on the basis of key words and truncation. This approach on the one hand produces a lot of erroneous "hits" such as finding "boil" when searching for the string "oil". On the other hand it misses pieces of relevant information if the term is expresed by synonyms or hidden in syntactic descriptions. Intelligent information retrieval needs a complex grammatical analysis of the text, and it seems only logical to use the linguistic base of an existing natural language system. However, if one attempts to extend the analysis of a machine translation system the purely translation-directed approach will not suffice. The focus of information retrieval is the concept, not the word. And so it is to be expected that transfer system lexicons now used for machine translation will migrate towards a concept orientation. The process will be slow since large numbers of lexemes must be reclassified, and certainly none of the older, less sophisticated MT systems will be able to follow the route. As a benefit to machine translation, such a powerful add-on semantic component might also supply a wider range of linguistic features to improve translation quality. Some transfer problems can only be solved by the incorporation of a multi-levelled knowledge base. (*cf.* Schneider 89).

6. THE PLACE OF TERMINOLOGY WORK IN BUILDING SYSTEM LEXICONS

Even if we include phraseology in the scope of translation-oriented terminology work, it seems that terminology work does not have quite the same status in the context of machine translation as it does in conventional translation.

For one thing, general vocabulary has to be considered as well as specialized terminology. Secondly, in all lexicon entries, even in those of terminological content, the linguistic, non-terminological data play a decisive role. Thirdly, one can assume that in translation bureaus who opt for machine translation a lot of the necessary terminology has been collected already. Alternatively, for some MT systems the vendors may offer precoded lexicon modules for selected subject fields so that the end user can save the time of terminology research and coding.

It is understood that the compilation of defined terms and terminologies and their equivalences across two languages is the prerequisite for the building of a system lexicon. It is procedurally important that the validation of terminology precedes the generation of lexicon entries, or the work may suffer. The building of a system lexicon requires so much novel, sometimes seemingly strange steps that one should not attempt to do terminology work at the same time. Both tasks are best kept separate[6].

7. LEXICAL WORK IN A CHANGING TRANSLATION WORLD

In the past one could speak of a well-structured translation bureau if it had a central terminology service separated from the translation sections performing the productive

translation work. Today it has become apparent that many translation bureaus will eventually need another autonomous functional unit which prepares and supervises the use of a machine translation system. The tasks will — besides the usual logistics involved in handling machine-readable texts from various sources — mainly center around lexical work. While the system grammar is supplied, lexicon update will always be necessary. This function may evolve from an existing terminology service. If such a service does not yet exist, it might be beneficial to introduce both functions in the same organizational unit but it must be ensured that sufficient room is made for autonomous terminology work. Such a central function might prove more efficient and economical than spreading identical tasks across large groups of non-specialists.

The main task in creating entries for a system lexicon is to supply the necessary linguistic coding. This presupposes a good understanding of the interaction between grammar and lexicon and the principles of terminology work as well as some knowledge of data processing. For lack of a better word the person responsible for lexicon maintenance and update might be called a "system lexicographer". In the past, no field of study had led to the necessary qualifications. But with the inevitable advent of machine translation it might be worthwhile to integrate the subject into translator training programs. A half year course containing the basics of working with data processing systems, of terminology work, linguistics and the efficient use of machine translation may greatly enhance chances of employment for translators.

NOTES:

1. *Cf.* I. Hohnhold (1988): "Der terminologische Eintrag und seine Terminologie", Mitteilungsblatt für Dolmetscher und Übersetzer 34, 1988, nº 5.
2. An indication of the time of the research, of relevant source material and of the methodology of the terminology work can aid the user in assessing the quality of the records. References to pertinent literature which may be a good starting point for further research on the part of the user could trigger a translator's latent desire to start his own activities in compiling sets of terminology.
3. For example, a terminology data base may be queried. Such a data base constitutes the ideal store for terminological data, no matter the degree of differentiation or scope. From a data base, any informational item listed in the record may be queried on its own or in connection with other items. For the technical translator, the combination of data bases plus specialized dictionaries derived from them provide the most extensive lexicographic aid. If in the past, terminology data bases only existed on main frames, by now several data base systems have been made available on PCs (*e.g.* Term-PC by Siemens).
4. The METAL system is described in detail in: T. Schneider, "The METAL system. Status 1989" in Ch. Rohrer, ed. MT Summit II, Frankfurt: DGD 1989, pp. 165-173, and in T. Schneider, "State of the Art in West Germany" in Benefits of Computer Assisted Translation to Information Managers and End-Users, AGARD Lecture Series 171, Neuilly-sur-Seine: AGARD 1990, pp. 8-1 — 8-12.
5. For details see T. Schneider, "Problems of Machine Translation and Semantic Knowledge" in L. McCrank, ed. Databases in the Humanities and Social Sciences 4. Medford: Learned Information Inc. 1989, pp. 479-586.
6. The different content of terminological records and system lexicon entries may lead to a lack of precision in the terminology work if both have to be generated side by side, especially when productive tasks have to be dealt with under time pressure.