PRACTICAL EXPERIENCE OF MACHINE TRANSLATION
V. Lawson (ed.)
149
North-Holland Publishing Company / © ASLIB, 1982

THE PIVOTAL ROLE OF THE VARIOUS DICTIONARIES IN AN MT SYSTEM

Francis E. Knowles

(Department of Modern Languages,
University of Aston in Birmingham,
BIRMINGHAM, GB)

Long before the advent even of rudimentary MT systems linguists (and philosophers) were attempting to fathom out and describe the relationship of the "potential meaning(s)" of a word or phrase as a dictionary entry and its "actual meaning" by reason of occurrence in text. The implementation of MT systems requires an adequate formalisation of this relationship and although a semanto-syntactic environment is frequently sufficient to enable the choice of a correct lexical item this paper shows that, however powerful particular parsing techniques may he, a crucial role attaches to the dictionary component of any MT system.

Linguistics does not provide its students with an unambiguous answer as to how linguistic meaning, an inalienable part of linguistic systems, is to be treated or investigated. Linguistics does not formulate a single consensus about meaning, it leaves the door open to many approaches of which two opposing ones are worthy of note in this immediate context. These "rival" approaches compete as to the object of semantic analysis — where is meaning to be sought? In Saussurean terms, is this object a unit of "langue" - such as the "word" - or is it in the domain of "parole" - in the form of an utterance? Does the "word" have a semantic autonomy or does it only acquire meaning in association with a context composed of other "words"? This dichotomy of approach - which stares MT researchers in the face from the word "go" - is tantamount to a choice between a lexicocentric philosophy and a textocentric philosophy. It can no longer, of course, be seriously suggested that these two approaches are self-excluding - although the earliest MT attempts were entirely lexicocentric, even to the point of arraying lexical alternatives for the human reader to select from! Some more recent MT systems have - in effect - gone to the other extreme - by falling prey to the natural temptation of allowing the process of scanning actual texts to be the major factor and philosophy in the elaboration of machine dictionaries, the main problem being that the entries in these dictionaries were actually segments from utterances!

I suggest that in order to develop a reliable lexical data base comprising either a series of sub-dictionaries or a fully invertible network it is necessary to have available information relating to the total set of potential meanings of "sense units" and to use (con)textual information to filter out the "noise" and leave either an actual meaning or, if circumstances force it, a virtual meaning. In the easiest case, sense units can be isolated iconically, by straight matching, and the "burden" of semantic analysis and interpretation can be passed across to the human reader to whom this function properly belongs. The longer the matched segment is, the less is any probability of ultimate error. In many of these cases, multi-word units - that is, semantically "atomic" units consisting of more than one orthographic word - are involved and the process of isolating them is akin to routine syntactic analysis, although the lexical process is one of syntagmatics rather than syntax. "He has been sent to Coventry" is an example of a candidate for an iconic match, presumably in an idioms dictionary structured to give a reasonable equivalent in a particular target language. Incidentally, "He is in Coventry" is potentially a much more difficult problem as it contains only one nucleic word, as opposed to two - "sent" and "Coventry" within a critical span. Obviously, much depends on context but that statement will need to be amplified below. If this method fails - and it often does - then syntactic features, contextual features, paradigmatic features appertaining to literally - crucial words must, of course, be used in attempts to cut down maximally the range of alternatives, the least plausible being discarded first in the progression to a unique hypothesis with a high probability that is nevertheless smaller than unity, the more so whenever extra-linguistic factors are concerned.

One frequently suggested and practised method is to formalise a function to compound a sentence's syntactic analysis, semantic interpretation and its "setting". However, the "setting" parameter is, as yet, a blunt instrument which tends either to confirm more than one interpretation or deny the possibility of any! The bluntness of this instrument is, of course, directly related to the failure to penetrate deeply enough into the intricacies of "real-world" settings. This criticism is meant to be kindly, however, given the immensity of the problem and the analogous difficulty, in "ordinary" lexicographic practice, of deciding how much encyclopaedic information to include in a dictionary. As against "He is in Coventry" let us consider "He is in Wormwood Scrubs". In order to translate - usually by explicit amplification in human translation - the machine dictionary must contain an entry "Wormwood Scrubs" and a gloss such as "a well-known British prison".

The problem of the setting is a big one because it often occurs that the correlates of a setting are interrupted by sentence boundaries. That does not worry those engaged in the non-computerised study of text linguistics because it is text linguistics rather than sentence linguistics. In MT work it has often been felt reasonable to supply - in the interests of disambiguation - topic parameters, such as: "this passage is about penal reform", which steer, inter alia, dictionary/glossary selection and can successfully bypass genuine disambiguation by freezing out non-viable settings. Nevertheless, if the topical glossary parameter is given as "mechanical engineering" it would not ease on or tease out a correct translation for a text snatch I found recently: "Fitting the wrong sleeve would be putting a major spanner in the works"! The equivalent idiom in German, for instance, would be: "jemandem einen Knüppel zwischen die Beine werfen", literally "to throw a cudgel between someone's legs" - which is, I suppose, the same as putting a spoke in someone's wheel etc.etc!

To return to the prior point, it can be said that the search for methods of building up a "semantic component" on the basis of formal hierarchies or of a conceptual calculus for semantic universals or of a setting, script or menu - call it what you will -

has led to a focussing on the complete utterance and to a shifting away of interest from the individual units comprising the utterance. This in turn has demoted "langue" and lexicology and has promoted "deep semantics" and semiology. Yet, the semantic interpretation of many utterances is impossible without a detailed analysis of its constituent parts - often mere orthographic words - and that cannot be done without recourse to extralinguistic factors which largely help to determine the semantic units. To deny this is to postulate linguistics without language.

According to Luk'janova (Luk'janova), the design of ideal linguistic software and data base systems is impeded by difficulties of two kinds: firstly, by the problems attendant upon the formalisation of - semantically - unformalised textual information and, secondly, by the subtleties required in terms of program and data base coding for real systems. Leaving this second point until later we address ourselves to the first by recalling that natural language is an open-ended system for information transfer and it relies on the fuzzy sets which, par excellence, characterise linguistic data. A computer memory can store most easily a closed—off, "shorthand" description of natural language - no-one has, to my knowledge, implemented an extensive fuzzy-set system complete with all the statistical information it requires. Without methods of this sort it will not be possible to develop algorithms for computerised "associative thinking". There are other factors of serious import. Any text generated, created by the interaction of semantic information and a particular linguistic system possesses a statistical structure which is both a combination of universals - such as redundancy level - and of features characteristic of its actual language, style, register and of its individual author. In other words, we have the antinomy of "langue" and idiolect, whereas the formal grammar would "assume" that idiolect features had been subsumed without vestige. If such features were present - and we have stated that they always are - they would be by definition "deviant".

Furthermore, in order to function effectively, a man-machine system, such as a MT system, presupposes the residence in core of a model of real-world concepts and objects that is closely correlated to the "models" operated by the human users of the system. This requirement is impossible to meet, not only in practical terms but also in principle, at least given present insights into the nature of the problem. In my view, MT systems designers must reconcile themselves to the need to build into their design the informed human reader of MT output, one of whose tasks it is to reconstitute the totality of messages, given parts that add up to less than the whole. High-quality fully automatic machine translation means keeping this shortfall of information as small as possible. As to what "small" means, experiments on the readability of texts presented to language learners have shown (Piotrovskij - 1973) that readers extracting 80% of the actual information in such texts usually have no difficulty in closing the information gap satisfactorily. However, the use of the word "usually" reminds us that we are dealing not with a deterministic but with a stochastic state of affairs. There is, of course, the problem of the proportioning of the information extracted between lexical, syntactic and thematic categories.

As our topic is lexicography, let us now focus our attention on lexis as such. A word cannot help, in people's minds, forming conceptual relationships with other words. One particular type of relationship is associativity and its concurrent notion of "substitutability". A simple example of this would be the (incomplete) set of words: (pie, flan, bun, cake, tart, scone). The elements of the set are similar yet they are in contrast, and all "line up", vying for insertion into appropriate utterances. This relationship is referred to as paradigmatic and is clearly distinct from the syntagmatic relationship a word assumes when it is arrayed - to use a textual image - side by side with "neighbours" in a linear fashion. It is the duty of lexicographers - and that includes the elaborators of MT lexical data bases - to "capture" all the paradigmatic relationships of a word and as many of the superordinate set-to-set relationships as are also possible. What must also be "captured" are all those syntagmatic relationships that are not purely volatile, where "volatile" means an association not known to be statistically significant. Hence it is a search for all multi-word units that exist prior to the sentence and outside the sentence. This is a daunting task but a tantalising one, especially with regard to polysemous words.

Where the syntagmatic relationships are not volatile but are "weighted" statistically then the lexicographer - either within a MT context or without - has a task that gets harder as the statistical weighting gets lighter. He includes in his general dictionary or in his idioms dictionary items which have a high degree of bonding, or of predictability, once "commenced": such items include all proverbs - such as: "better late then never" or "live and learn" - idioms with a motivated meaning - such as: "to let off steam" or "to hit below the belt" - phraseological concretions not analysable into constituent parts - such as: "to cut off with a shilling" or "to win one's spurs" - collocations - such as: "Pyrrhic victory" or "curate's egg" - and multi-word units - such as: "fair and square" or "once and for all", plus many, many technical terms such as "yellow fever" or "small intestine". "Yellow fever" and "small intestine" have, incidentally, one-word equivalents in German, being "Gelbfieber" and "Dünndarm", respectively. The lexicographer's chief difficulty is drawing a line between specific lexicographical units and "algebraic" models embedded into a more overtly syntactic concept and known as valency or case frames. Case frames are useful, among other things, for making a correct lexical choice during the generation of a target text. An example of this would be the translation into German of: "He finished his cake and she finished her wine". Analysis would show the first "finish" verb had an argument denoting solid food, whereas in the second use it was a case of liquid food. The following translation could then be produced: "Er ass seinen Kuchen auf und sie trank ihren Wein aus". A secondary but associated difficulty is the demarcation between syntagmatics proper and microcontext. (Macrocontext goes, of course, beyond sentence boundaries and is unfortunately "out of court" for that reason!) Translation difficulties would therefore result in a "loose" system with an utterance such as: "Cheap wine is flooding the market". The verb would be taken as the governor of the case frame, which would then programatically ascertain that "wine" is a liquid and "market" is a physical location - which could, conceivably, be flooded! Similar difficulties could result with the lexical collocation - or syntagma - "wine lake" unless that particular syntagma had been "lexicalised", that is, entered as such into the dictionary. More of this later.

Even the most ardent generativists sometimes appear to stop short in their tracks at the suspicion that the "lexical component" of their grammar systems is virtually static - they normally recover their composure after deciding that their chief and almost exclusive focus of interest is centred on the nature of the dynamic and volatile relationships between "data objects" representing sense units and appearing - predominantly, it is often supposed - in the guise of the orthographic words that comprise running text. Yet experimentation with information-theoretic methods has shown (Piotrovskij - 1973) - and confirmed Luk'janova's "information-pragmatic" results - that at least 80% of any text's information is embedded in its lexis and phraseology, this latter term being taken in its lexicographical sense. The point here - and I do not wish to labour it - is this: however powerful and sophisticated a particular set of parsing algorithms may be, the successful analysis of a text into a correct meaning representation is only possible if a commensurately "powerful" and sophisticated lexical data base is wedded to the dynamic modules of the MT system.

The use above of the phrase "meaning representation" may provoke the comment: "don't beat about the bush - we're talking about semantics!" My rejoinder to such a remark is to agree with it, whilst pointing out that we are really talking about a whole range of semantic sub-systems, each of them geared to capturing an individual type of meaning.

It is instructive for MT systems designers to ponder on Leech's convincing and comprehensive account of meaning (Leech), which he splits up into seven types. Let us review this account, musing on its implications for MT lexicography and semanto-syntactic analysis as we do so. Starting on familiar ground, we have "conceptual meaning" which can be and is codified satisfactorily most of the time - the traditional dictionary's job is to explicate conceptual meanings and the methods used for this are preferably those of componential analysis and the contrastive definitions which emerge from this process. A "classical" definition of the definition process is given by Rubinstein and Weaver: "Definition is process of placing a word (or Term) into a family (or Genus) and then separating the word from the other members of the family by showing the difference (or Differentia). Example: A catalytic agent (Term) is a chemical (Genus) which quickens the reaction of other chemicals, but is itself unchanged (Differentia)." The measure of success in this enterprise is determined by the judicious and difficult choice of a set of distinctive features capable of accurate and non-redundant application. An example of the technique would be the partial specification of the word "vegetable" as: +plant, +herbaceous, +food. This specification of this meaning - the meaning: "dull/inactive/brain-damaged human" is not catered for - could be tightened up by the addition of further feature values. The traditional dictionary, of course, tries to specify meaning in a similar way but discursively - the binary feature "shorthand" is, however, a highly formal method which makes it conveniently amenable to computerisation. It goes without saying that no-one anywhere has yet derived feature sets for really large-scale use, although sets of 96 features for verbs and 43 for nouns are known to be in use in one system. (Bilan) The technique is used for disambiguation purposes and has performed reasonably well in some systems which invoke comparisons between the feature profiles of doubtful items, searching for matches to confirm a unique interpretation.

A second type of meaning is connotative meaning which goes beyond conceptual meaning in that, as the term implies, it connotes rather than denotes. This connotation process is "sited" in the area of extra-linguistic experience and of the elusive socio-cultural consciousness of individuals, groups, communities and nations. It also affects the formation and formulation of political attitudes. All of this implies the occurrence of potentially serious translation problems. In translation of whatever sort - MT or HT - these problems are intensified whenever the realia of the source language cannot really be mapped onto equivalent target language realia. There is no satisfactory way of translating "floating voter" into Russian, for instance, and there are always going to be mismatched connotations inherent in the translation of "alcoholic drinks" from English into Arabic, similarly. However, these problems are minimised whenever the author of documents to be translated is skilful and ponders his words carefully. If, however, the author does not know at the time of writing that his work is to be translated, or if he is ignorant or dismissive of connotative aspects of meaning then his translator has an ethical problem to solve before he begins, i.e. how much should he accommodate. An MT system has not an ethical problem, just a mechanical one, which it can occasionally - only occasionally - give the appearance of having solved if the MT dictionary system contains good idiom, metaphor and simile equivalences or if, vacuously, lexical substitution leaves the connotative position unchanged, so to speak.

The third type of meaning is stylistic meaning, where "stylistic" refers to features such as individuality, dialect, chronological setting, mode of discourse, text type and subject field. Most of these features are, operationally speaking, irrelevant to MT but there is a crucial need to specify and model as closely as possible the features of text type and subject area. Of these, the former is manifest predominantly on the syntactic level, with ramifications from the sentence level upwards, via paragraphs to complete texts. It always helps to recall the Latin etymology of the word "text" - "textus", meaning "woven" and hinting metaphorically at a carefully woven "fabric", a fine texture, in fact. To take an extreme example of text type, however, the English preposition "from" translates into the German noun "Absender", or its standard abbreviation "Abs.", if the text type is "the back of an envelope", so to speak. Although therefore a MT system's control over and handling of the "text type " parameter are more directly correlated with the adequacy of its grammatical models rather than with its lexical data base, the "subject area" parameter is, conversely, largely catered for and underwritten by the amount of care put into the establishment of the terminological glossaries characterising not just broad-focus but preferably narrow-focus subject areas.

Affective meaning, Leech's fourth category, being concerned with the "revelation" of the author's own attitudes to situations or ideas, is described by him as parasitical in the sense that its "system" is more or less totally merged and submerged with the conceptual, connotative and stylistic systems.

Reflected meaning occurs whenever a lexeme has more than one conceptual meaning and a reader's response is evoked not only for the actual meaning but also for the other(s). Many of the SYSTRAN mistranslations which I quoted to this ASLIB audience three years ago were case of mistaken identity and their reflected consequences. (Snell) This feature can, of course, be employed intentionally - it is the basis of most jokes, adverts and slogans - such as the very topical one: "Unemployment is not working!"

Collocative meaning is that portion of total meaning which derives from a particular lexeme's associations by way of co-occurrence with other words. Particular "slots" for this type of meaning are qualifying adjectives and either verbal complements or the verbal operand for a particular substantive. Collocability is lexically conditioned on the level of individual lexemes and is therefore a major task for both human and machine lexicography. It assumes particular importance at the synthesis stage of MT when, presumably, a semantic representation in interlingual or formal-logic notation has, among other tasks, to be driven upwards towards a surface representation. The most interesting work in this aspect of MT dictionaries has, in my view, been done by Melchuk and his colleagues who defined approximately fifty abstract lexical functions to solve the collocation and other problems. Simple examples of these functions would be:

a function "Verus", denoting "correctness, truth, appropriateness" -

<u>lex</u>	<u>verus(lex</u>)
guess	correct
sentence	just
pride	legitimate
comment	apposite
suspicions	well-founded
behaviour	model
citizen	loyal
prediction	prophetic

or, a function "Operand" linking the name of the first "actant" in the role of the subject within the name of the "situation" in the role of the object -

<u>lex</u>	<u>operand(lex</u>)
steps	to take
war	to wage
assistance	to render
treaty	to conclude
contact	to maintain
investigation	to conduct
deal	to negotiate

Should the name of this latter function be "Cliché"?

Leech's last type of meaning is thematic meaning which focuses on the role played by word order and emphasis in the "message" as a whole. Although closely bound with text type and with overtly syntactic matters, the success of an MT system in handling this aspect of meaning depends in no small measure on the extent to which the lexical data base contains the information necessary to successfully convert individual lexical units from one part of speech to another prior to the re-assembly of thematic meaning in a suitable target language syntactic formulation. A standard situation would be the English "She arrived two hours before I left" translated into German as: "Sie kam zwei Stunden vor meiner Abfahrt an", literally: "she arrived two hours before my departure". Taking the case of the reverse translation of this sentence, a similar suggestion is that the decipherment of a deverbal noun should not be accomplished by a direct dictionary entry but rather by "indirection" to the appropriate verb - all in the context of the frequent syntactic phenomenon of nominalisation. Otherwise knotty translation problems such as: "he is a heavy eater" can be solved by this method. Similar things could be said about deadjectlval nouns, denominal qualitative adjectives or adjective/adverb flip-flop systems, and this opens the door to a much more comprehensive treatment of word-derivational morphology and etymologically-based nets.

I should like to quote as an example of a automated lexical data base the Soviet ASNTI/BOLID system, the design parameters of which extend beyond the needs of MT and cover a broader focus which takes in tasks - performed on-line - such as fact and document retrieval, bibliographical searches, automatic abstracting and indexing, and, not least, the processing of managerial documentation to expedite executive decision-making. (Luk'janova) It has been found that in spite of the wealth and variety of tasks to be performed it is possible, desirable even, to utilise one linguistic data base: this is a system consisting of Russian and foreign glossaries in which each of the entries represents a lexical unit to which is appended information relating to the grammatical, semantic, phraseological, terminological and thesaural status of the entry. Each of the automated dictionaries is divided into two sections. The grammatical information referred to would be values from a set of lexico-grammatical codes determining the morphosyntactic categories of the lexical unit's forms. The semantic information is represented by code values for the lexical unit's semantic classes within a specified sublanguage. The phraseological information section contains pointers to the idioms dictionary. The terminology section contains the codes of the sublanguages in which the given lexical unit is used and also defines the unit's terminological (or even descriptor) status. The thesaural section carries values denoting genus-species relationships and other associative links.

Each automated dictionary is arbitrarily divided into two sections: a general-purpose section and a terminological section. The working premiss is that the general-purpose section and the "general" idioms dictionary are constant, whereas the terminology sections vary in accordance with the texts being processed. The general-purpose dictionary was elaborated statistically outside a MT framework and contains items from all subjectively "general" sources which have been culled from frequency dictionaries. The "admission ticket" of lexical units to the general-purpose dictionary is granted after it has been established by a number of criteria that there is a high degree of occurrence correlation between their sources and that their global distribution is smooth and closely fits appropriate theoretical curves. Words characteristic of specific sublanguages or exhibiting different semantic characteristics in different sublanguages are assigned to appropriate topic glossaries. Apart from a small (50 items) high-frequency dictionary which is maintained in core the lexical data base is kept on spinning store and is played through the system as required.

This is a suitable juncture at which to go into some quantitative aspects of the BOLID system. As regards the sublanguage glossaries just mentioned, the computer engineering glossary for English was elaborated from texts totalling 200,000 tokens and yielding 13,160 types. Its Russian equivalent "started life" as a corpus of texts 250,000 words long, from which a glossary of some 10,520 entries emerged. In both cases the entry or type count includes so—called idioms, which are generally stable collocations consisting of two or three orthographic words. In terms of comparisons with "human" dictionaries, Webster's contains close on half a million entries. Modern thesauri often contain over ten thousand descriptors and keywords, themselves sieved from technical discourse most often using lexemes not found in dictionaries such as Webster. On this basis it is possible to hazard the guess that English lexis amounts to perhaps as many as a million lexical units. Estimates for Russian run to over 800,000 lexical units, incidentally.

There is one significant linguistic contrast between Russian and English, however, and this relates to the typological differences between these two languages and this, in its turn, leads to differences in the machine implementations of the Russian and English lexical data bases. English is predominantly an isolating language with only a residual inflectional system. Russian is a highly inflectional language but is not totally devoid of function words or of analytical grammatical forms. In terms of the data compression necessary for efficient lexical data bases the English dictionary in BOLID is a dictionary of word forms whereas the Russian dictionary is composed of so-called "machine fragments", that is, truncated stems which do not necessarily correspond to morphemes, either existing or putative! To these fragments can then be agglutinated their grammatical endings. That sounds easy but there are 31 declensional types for Russian adjectives, the paradigm of which possesses 32 fields. The corresponding numerical indices for substantives are 190 declensions and 12 paradigm fields; for verbs the figures are 328 conjugations and 33 paradigm fields. Taking these data and other factors, undiscussed, into account we are confronted with the nightmare of a Russian-language lexical data base of over 500 million symbols in "length"!

I have not dwelt in this paper on the properly logistic structures of MT dictionaries, preferring for today to concentrate attention on the prior problem of their design. Nevertheless, it is appropriate to turn our attention briefly to the relevant logistic aspects of MT dictionary implementation. One operational point to note straight away is that the phrase "lexical data base" mirrors better the required situation, given the flexibility and power of data base management systems. A number of special facilities peculiar to machine lexicography are required in addition. These include - in EUROTRA thinking, for instance (EUROTRA) - facilities for:

1) expanding entries from their "tight" internal formulation to a "human reading" format;
2) defining a set of morphological and other classes which automatically prompt the coder;
3) automatically generating regular word forms;

4) copying and deleting entries in part or in toto;
5) inserting entries whilst maintaining the security of already existing entries;
6) editing entries by parameter or after inspection;
7) automatically verifying the presence of "counterparts" as between the analysis and transfer sections of the lexical data base;
8) automatically cross-checking between headwords standing alone and occurring as part of multi-word expressions;
9) maintaining an "air-lock" system to prevent patches being added to the master data base without managerial permission;
10) on-line up-dating/editing for privileged users.

It may have struck some members of the audience as odd that I did not mention any cross-checking between the entries of the analysis and generation dictionaries in EUROTRA. This was due to the following important point: according to the technical specifications of the system, the entries in the EUROTRA lexical data base consist of three parts - an identifier which invokes the dictionary rule in question and which may serve as an access mechanism; a representation which consists of a string of tests to be carried out on the isolated items of the text under analysis; and a set of values to be assigned if and when a particular test is successful. The important implicit concept is that dictionaries organised along these lines are reversible, a rough analogy being that a given definition may become a headword and its "old" headword may become its "new" definition and this whether the original mapping was simple or complex.

If the frequency characteristics of lexical units are to be taken into serious account — and I do not see how they can be wilfully ignored in a large-scale system - then, in my view, the MT system should incorporate a module - to be flipped on or off - to automatically, i.e. during routine translation batches, update frequency statistics and thus fine-tune all-important knowledge relating to the quantitative behaviour of all possible parameters in the lexical data base. Bearing in mind my earlier remarks about the danger of allowing commercial MT systems to adopt such a procedure ab initio, I suggest that all MT systems should have to accumulate an impartially stipulated logged total of words processed before being allowed to lay claim to genuine commercial viability. This is particularly important in the case of "restricted text type" systems or of systems likely to use extensive in-house nomenclature or other types of constrained terminology.

Particular importance attaches to principles established to treat "not found" items during commercial runs or even pre-commercial pilot runs. Often no principles are observed at all. The obvious danger is that of adding items to the lexical data base piecemeal and of thereby introducing inconsistencies. An analogy to this situation is appropriate from the world of mathematics. The danger is dealing exclusively with arithmetic and failing to extract the algebra. The same sort of point is made, but in a different way and in a different context, by King and her colleagues who state: "It should also be emphasised that there is no clear borderline between dictionary entries and grammar rules: they have the same basic structure. Any difference between them is based on the linguistic model used, in that dictionary entries describe the linguistic behaviour of individual lexical units where grammar rules describe more general linguistic phenomena". (EUROTRA) There is one further subtlety: knowing the difficulties of establishing an algebra that is general enough, either actually or potentially, to justify the effort of deriving it and specifying a level of criticality for it, the temptation is to add in "straight" to the lexical data base items which are, in fact, eminently amenable to an "algebraic" analysis. This anti-stratagem is known as "lexicalisation" - it avoids the very issues with which the MT lexicographer should be most vitally concerned, not least in the matter of his job satisfaction! The claim that "it works" is often equivalent to procrastinating and obfuscating a genuine lexical systems approach. It must never be the default approach. If it is, then, however apparently impressive "performance" may be, that performance is - if you will forgive the "word" - based on a "lexi-con-trick" !

REFERENCES AND BIBLIOGRAPHY

Ju.D. Apresjan: Leksičeskaja semantika
("Nauka", Moscow, 1974)

E. Agricola: Semantische Relationen im Text und im System
(Mouton, The Hague, 1972)

V.N. Bilan et al.: Metody avtomatičeskogo analiza i sinteza teksta
(MGPIIJa, Minsk, 1977)

I. Brand et al.: Automatische Sprachübersetzung, I-III,
(Akademie-Verlag, Berlin, 1967/72/76)

H. Bruderer: Handbuch der maschinellen und maschinen-unterstützten Sprachübersetzung
(Verlag Dokumentation, Munich, 1978)

N. Cercone/R. Mercer: Design of lexicons in some natural language systems
("ALLC Journal", Vol.I, 1980, pp.37-54)

S.M. Deen: Fundamentals of data base systems
(Macmillan, London, 1977)

A. Duff: The third language - recurrent problems of translation into English
(Pergamon, Oxford, 1981)

H-J. Diller/J. Kornelius: Linguistische Probleme der Übersetzung
(Niemeyer, Tübingen, 1978)

A. Duff: The third language - recurrent problems of translation into English
(Pergamon, Oxford, 1981)

H. Eggers: Maschinelle Übersetzung, Lexikographie und Analyse, Vols.I-II
(Saarland University, Saarbrücken, 1980)

EUROTRA: Technical Specification (EUROTRA/160/80)
(CEC, Luxembourg, 1980, /restricted/)

P. Francois: La banque de données terminologiques EURODICAUTOM
("La maison du Dictionnnaire", Paris, 1977)

K-H. Freigang et al.: Der Stand der Forschung auf dem Gebiet der maschinellen Übersetzung
(Saarland University, Saarbrücken, 1979)

V.G. Gak: Sopostavitel'naja leksikologija
("Meždunarodnye otnošenija", Moscow, 1977)

C.C. Greenfield/D. Serain: La traduction assistée par ordinateur: des banques de terminologie aux systèmes interactifs de traduction
(IRIA, La Chesnay, 1977)

G. Guckler: Zweisprachiges Wörterbuch für angenäherte operationelle Analyse semantischer Entsprechungen mittels EDV
(Narr Verlag, Tübingen, 1975)

F. Güttinger: Zielsprache - Theorie und Technik des Übersetzens
(Manesse, Zurich, 1963)

R.R.K. Hartmann: Dictionaries and their users
(Exeter University, Exeter, 1979)

B. Henisz-Dostert: Machine translation
(Mouton, The Hague, 1980)

L. Hoffmann: Fachsprachen und Sprachstatistik
(Akademie-Verlag, Berlin, 1975)

F. Hundsnurscher: Neuere Methoden der Semantik
(Niemeyer, Tübingen, 1971)

W.J. Hutchins: Progress in documentation and machine-aided translation
("Journal of Documentation", 1978, pp. 119-159)

Ju.N. Karaulov: Obščaja i russkaja ideografija
("Nauka", Moscow, 1976)

F.E. Knowles: Recent Soviet work on computer techniques for representing natural language meaning
(in: MacCafferty/Gray)

V.N. Komissarov: Lingvistika perevoda
("Meždunarodnye otnošenija", Moscow, 1980)

G. Leech: Semantics
(Penguin Books, Harmondsworth, 1974)

W.P. Lehmann/R. Stachowitz: Machine translation in Western Europe
(in: Sebeok - Vol.IX)

A. Ljudskanov: Mensch und Maschine als Übersetzer
(Niemeyer, Halle, 1975)

E.M. Luk'janova: Informacionnaja baza avtomatičeskix slovarej
(in: Piotrovskij - 1980)

J. Lyons: Semantics, Vols.I-II,
(C.U.P., Cambridge, 1977)

M. MacCafferty/M. Gray: The analysis of meaning - Informatics-5
(ASLIB, London, 1979)

J. McNaught: Terminological data banks: a model for a British linguistic data bank (LDB)
("ASLIB Proceedings", Vol.33,
1981, pp. 297-308)

I.A. Mel'čuk: Opyt teorii linvističeskix modelej 'smysl-tekst'
("Nauka", Moscow, 1974)

I. Pinchuk: Scientific and technical translation
(Deutsch, London, 1977)

R.G. Piotrovskij et al.: Tekst, mašina, čelovek
("Nauka", Moscow, 1973)

R.G. Piotrovskij: Inženernaja lingvistika
("Nauka", Leningrad, 1979)

R.G. Piotrovskij et al.: Statistika reči i avtomatičeskij analiz teksta - 1980
("Nauka", Leningrad, 1980)

A.H. Roberts/M. Zarechnak: Mechanical translation
(in: Sebeok - Vol.VII)

S.L. Rubinstein/R.G. Weaver: Frameworks of exposition
(Holt, Rinehart and Winston, New York, 1966)

T. Sebeok: Current trends in linguistics
(Mouton, The Hague,
Vol. VII, 1974,
Vol. IX, 1972)

G. van Slype/I. Pigott: Description du système de traduction automatique SYSTRAN
(CEC, Luxembourg, 1979)

B.M. Snell: Translating and the computer
(North-Holland, Amsterdam, 1979)

H.L. Somers/ R.L. Johnson: PTOSYS: an interactive system for "understanding" texts using a dynamic strategy for creating and updating dictionary entries
(in: MacCafferty/Gray)

H.L. Somers: ISSCO*PTOSYS
(ISSCO, University of Geneva, 1980)

P. Toma: SYSTRAN - ein maschinelles Übersetzungssystem der dritten Generation
("Sprache und Datenverarbeitung", 1977, pp. 38-46)

H.M. Townley/ R.D. Gee: Thesaurus-making: grow your own word-stock
(Deutsch, London, 1980)

J-P. Vinay/B. Kallio: Bilingual lexicography and the computer
(Department of Linguistics, University of Victoria BC, 1971)

S. Vlaxov/S. Florin: Neperevodimoe v perevode
("Meždunarodnye otnošenija", Moscow, 1980)

G. Wahrig: Anleitung zur grammatisch-semantischen Beschreibung lexikalischer Einheiten
(Niemeyer, Tübingen, 1973)

D.L. Waltz: The state-of-the-art in natural language understanding
(unpublished paper, Illinois University, Urbana-Champaign, 1981)

A. Wierzbicka: Semantic primitives
(Athenäum, Frankfurt/M., 1972)

Y.A. Wilks: Grammar, meaning and the machine analysis of language
(RKP, London, 1972)

W. Wilss: Übersetzungswissenschaft - Probleme und Methoden
(Klett, Stuttgart, 1977)

T. Winograd: Understanding natural language
(E.U.P., Edinburgh, 1972)

E. Wüster: Die allgemeine Terminologielehre
("Linguistics", Vol.119, 1974, pp.61-104)

L. Zgusta: Manual of lexicography
(Mouton, The Hague, 1971)

A.K. Zholkovski/I.A. Melchuk: Semantic synthesis
("Systems Theory Research", Vol.19, 1970, pp. 179-243)

H. Zimmermann: Das Lexikon in der maschinellen Sprachanalyse
(Athenäum, Frankfurt/M., 1972)