[*Translating and the computer 12: Applying Technology to the Translation Process*. Proceedings of a conference … 8-9 November 1990, ed. Catriona Picken (London: Aslib, 1991)]

Machine translation of natural languages: the TOVNA MTS solution: a learning system

Presented by Ami Segal

TOVNA Translation Machines Ltd, Jerusalem

SUMMARY

Traditional methods of computer science and linguistics have failed to resolve the ambiguities of natural languages and the incompleteness of language grammar in order to produce accurate and meaningful machine translation of natural languages. Utilising a revolutionary approach, TOVNA MTS has solved this problem.

THE TOVNA MTS SOLUTION

Unlike other machine translation (MT) systems, TOVNA MTS learns from examples, continually improving with use, in order to produce consistently accurate and meaningful translations. Based on linguistics, mathematics, logic and artificial intelligence, TOVNA MTS combines rule-based and example-based algorithms to resolve ambiguities and grammar problems through a five-phase translation process. In effect, TOVNA MTS learns languages the same way people do, from examples, eventually translating in the style of its user.

A UNIVERSAL MACHINE

TOVNA MTS is both language and machine independent. Unlike other MT systems, where the need to rewrite the software makes it difficult and expensive to add new languages, TOVNA MTS uses the same software for all languages. Language data is kept completely separate from the software in the linguistic data base. The system's capabilities and user-interface are the same for all languages. TOVNA MTS can run on most commonly-available computers. The minimum operating requirements are a 32 bit processor work station, 16 MB of RAM, and 300 MB of disk space. TOVNA MTS is UNIX-based and written in the 'C' programming language.

THE TRANSLATION BUSINESS

Accurate and meaningful translation of documents, technical manuals, contracts, and other information products is essential to the conduct of business and government affairs throughout the world. With the global translation market now reaching approximately $25 billion per year, TOVNA MTS allows companies to substantially reduce the expense incurred during the pre-editing and post-editing phases of a translation project. Time-consuming and repetitive manual tasks are routinely but necessarily performed by translators and editors. TOVNA MTS dramatically improves the productivity of translators and editors, reducing pre-editing and post-editing translation time and cost, while ensuring consistent, accurate, and up-to-date translations.

BACKGROUND

Since the beginning of the computer era, machine translation (MT) of natural languages has been one of the most sought-after goals of computer science. However, it has proved to be one of the most elusive. This results from the fact that human languages are so complex and so different from one another that traditional methods of computer science and linguistics have proved incapable of extracting meaning from a source text and constructing a grammatically correct text which conveys the same meaning as the original source language text in the target language.

The complexity of natural languages can be summarised as follows:

a. Ambiguity, which leads to incorrect translation; and
b. Incomplete specification of grammar, which leads to incomplete translations.

Ambiguity

As an example of ambiguity at the simplest level, we may cite the word ‘work’, which can be used as either a noun or a verb. A more complex example of ambiguity is the sentence, ‘Fruit flies like bananas’, where ‘flies’ may be either noun or verb and ‘fruit’ may be either the subject noun or the noun modifying the following noun ‘flies’. The grammatical part of speech assumed by these words depends on the idea that the sentence is intended to convey, that is, the meaning of the sentence. In this case there are two distinct and unrelated alternative meanings, and there is no clue in the sentence as to which is correct.

In attempting to ‘understand’ the source text, an MT system must be able to analyse correctly (or parse) the structures of sentences in the text. An incorrectly analysed phrase will be incorrectly ‘understood’ and either not translated at all or incorrectly translated. For example, in the phrase, **‘the violet and pink shirts’,** both ‘violet’ and ‘pink’ are adjectives modifying the noun ‘shirts’. An unsophisticated MT system might analyse ‘violet’ as a noun (the flower) and attribute quite a different meaning to the text.

Grammar

It has proved virtually impossible to provide a computer with a complete and rigorous definition of the grammar of even the simplest natural language; the rules for constructing phrases and sentences in a language are incomplete in their specification of grammar.

The challenge

At the core of the issue is the question of how to address the enormous complexity of natural languages. The traditional approach has been to write complex algorithms to deal with complex language problems, but the innumerable rules of even the simplest language (and their innumerable exceptions) make for very complex algorithms indeed. If the method of dealing with a newly discovered shortcoming of an algorithm is to complicate the algorithm even further, then the software rapidly becomes a bewildering and confusing patchwork, impossible to understand or maintain, where each modification creates more problems than it fixes. Reflecting the language’s complexity in the algorithms did in fact lead to unwieldy, intractable and essentially unusable systems in all the MT projects undertaken prior to the National Academy of Science’s ALPAC report of 1966, a report which justifiably condemned this approach as unworkable, and which led to the abandonment of many MT projects which had adopted this approach.

The TOVNA MTS Solution

The TOVNA MTS solution uses a different approach, based on a recognition that the specification of a natural language as a finite set of rules is probably impossible. Therefore, an MT system must employ a combination of rule-based and example-based algorithms, and somehow find a reasonable way to overcome inevitable ambiguities. Using a combination of linguistics, mathematics, logic, and artificial intelligence, TOVNA MTS actually learns from examples and continues to improve with use, adding a human dimension to machine translation.

GENERAL PROBLEMS IN MACHINE TRANSLATION

1. Ambiguity leads to incorrect translations.

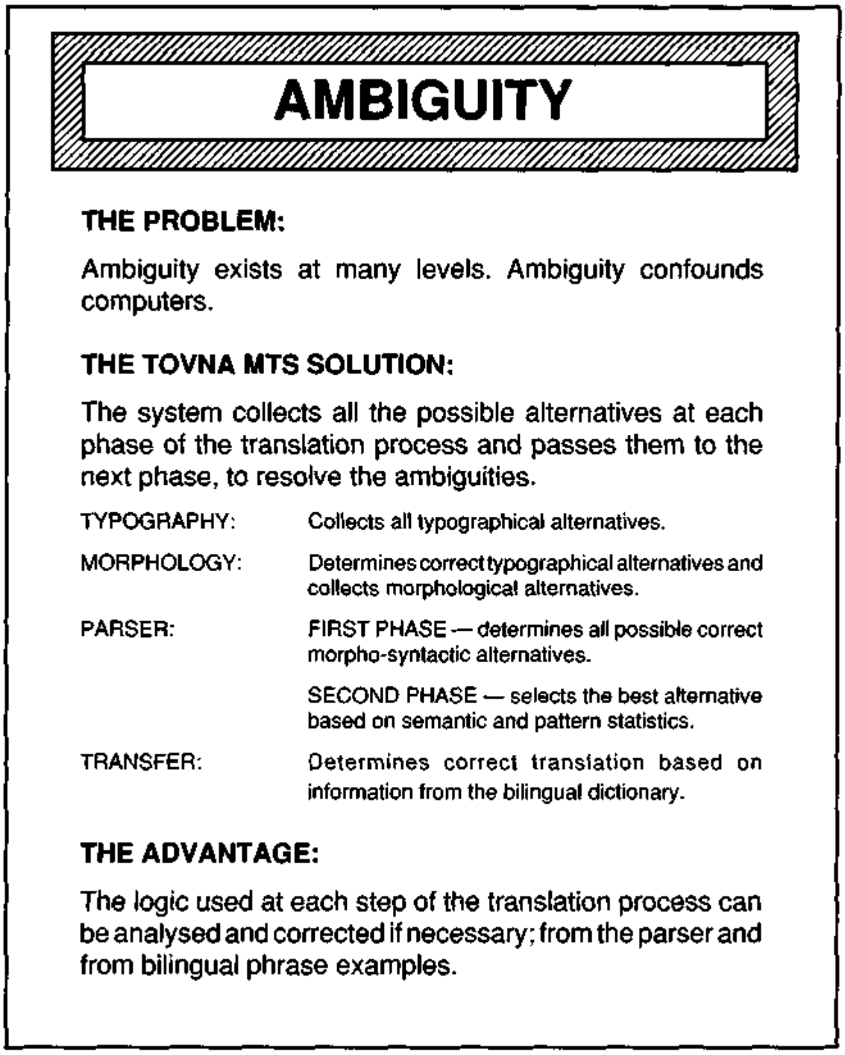
2. Incomplete grammar leads to incomplete translations.

3. Pre-editing and post-editing are time consuming and, in most cases, repetitive manual tasks.

4. It is very difficult to add new words to lexicons and dictionaries.

5. Additional languages are very difficult and expensive to add to an existing system.

6. Compatibility is not always available. Many systems do not integrate with existing computers, word processing software, and typesetting equipment.



AMBIGUITY

THE PROBLEM:

Ambiguity exists at many levels. Ambiguity confounds computers.

THE TOVNA MTS SOLUTION:

The system collects all the possible alternatives at each phase of the translation process and passes them to the next phase, to resolve the ambiguities.

TYPOGRAPHY: Collects all typographical alternatives.

MORPHOLOGY: Determines correct typographical alternatives and collects morphological alternatives.

PARSER: FIRST PHASE — determines all possible correct morpho-syntactic alternatives.

SECOND PHASE — selects the best alternative based on semantic and pattern statistics.

TRANSFER: Determines correct translation based on information from the bilingual dictionary.

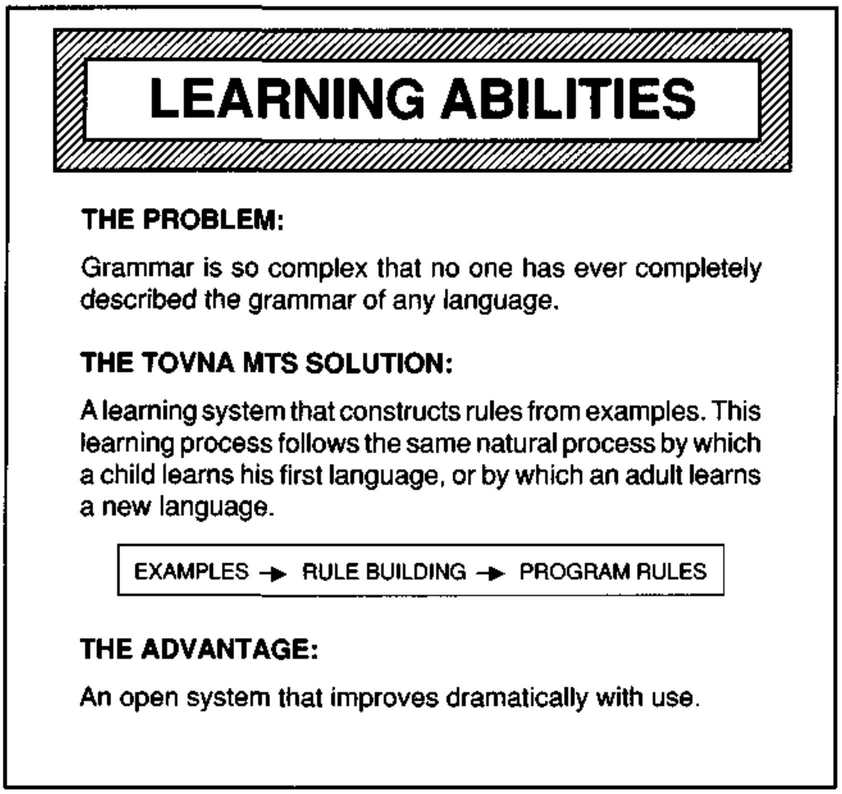
THE ADVANTAGE:

The logic used at each step of the translation process can be analysed and corrected if necessary; from the parser and from bilingual phrase examples.

THE TOVNA MTS SOLUTION

TOVNA MTS is a learning system which improves with use. The more it translates, the better its performance.

The TOVNA MTS solution to the problem of ambiguity (which leads to incorrect translation) is to discover, at each phase of the translation process, all the possible alternatives, and to pass them on to the next phase with the expectation that later phases will reject the incorrect alternatives.



LEARNING ABILITIES

THE PROBLEM:

Grammar is so complex that no one has ever completely described the grammar of any language.

THE TOVNA MTS SOLUTION:

A learning system that constructs rules from examples. This learning process follows the same natural process by which a child learns his first language, or by which an adult learns a new language.

EXAMPLES → RULE BUILDING → PROGRAM RULES

THE ADVANTAGE:

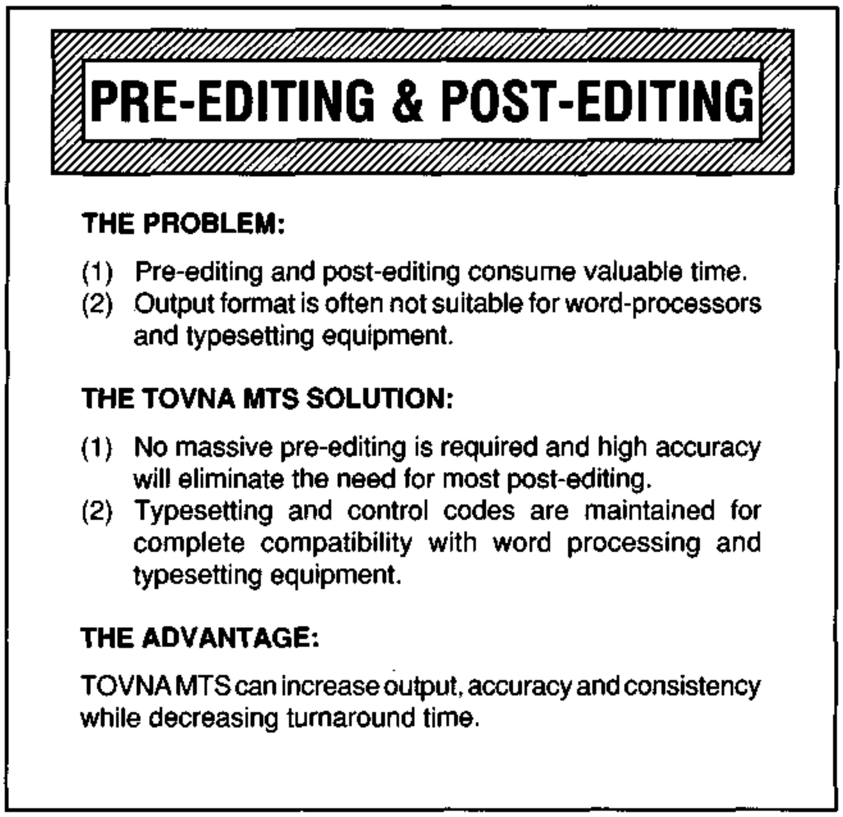
An open system that improves dramatically with use.

The TOVNA MTS solution to the problem of incomplete specification of grammar (which leads to incomplete translations) is its capacity to construct rules from examples. The user who ‘teaches’ TOVNA MTS a language’s grammar can do so by specifying an example. A user is never required to specify an algorithm.

This learning process follows the same natural process by which a child learns his first language, or by which an adult learns a new language. Because the system learns as it works, it will continue to improve. The more it works the more precise it becomes.

Although ambiguity and incomplete specification of grammar are crucial problems which must be solved by an MT system, they are hardly the only ones whose solution is critical to the success of the system. Other, less technical but still important issues which must be addressed are:

a. pre-editing and post-editing of text;
b. adding new words and phrases to dictionaries;
c. adding new languages to the system.



PRE-EDITING & POST-EDITING

THE PROBLEM:

(1) Pre-editing and post-editing consume valuable time.
(2) Output format is often not suitable for word-processors and typesetting equipment.

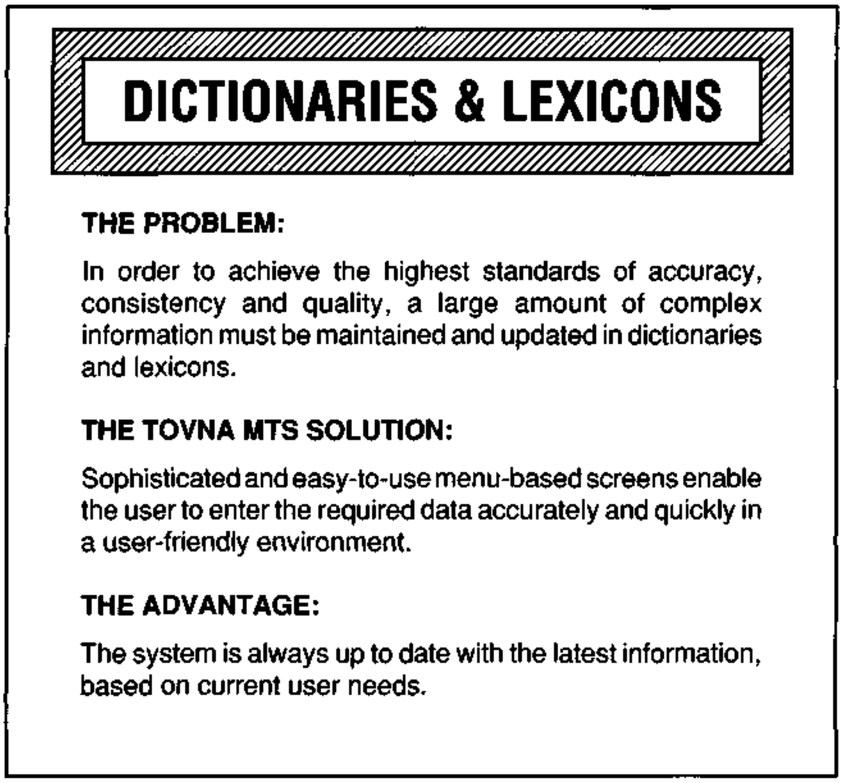
THE TOVNA MTS SOLUTION:

(1) No massive pre-editing is required and high accuracy will eliminate the need for most post-editing.
(2) Typesetting and control codes are maintained for complete compatibility with word processing and typesetting equipment.

THE ADVANTAGE:

TOVNA MTS can increase output, accuracy and consistency while decreasing turnaround time.

Pre-editing and post-editing of text consume valuable time. Complicating matters further, the output format is often not suitable for word processing and typesetting equipment. With TOVNA MTS, no massive pre-editing is required. A high degree of accuracy will eventually eliminate the need for most post-editing as the system improves its performance. Moreover, TOVNA MTS maintains typesetting and control codes for complete compatibility with word processors and typesetting equipment (existing and future).



DICTIONARIES & LEXICONS

THE PROBLEM:

In order to achieve the highest standards of accuracy, consistency and quality, a large amount of complex information must be maintained and updated in dictionaries and lexicons.

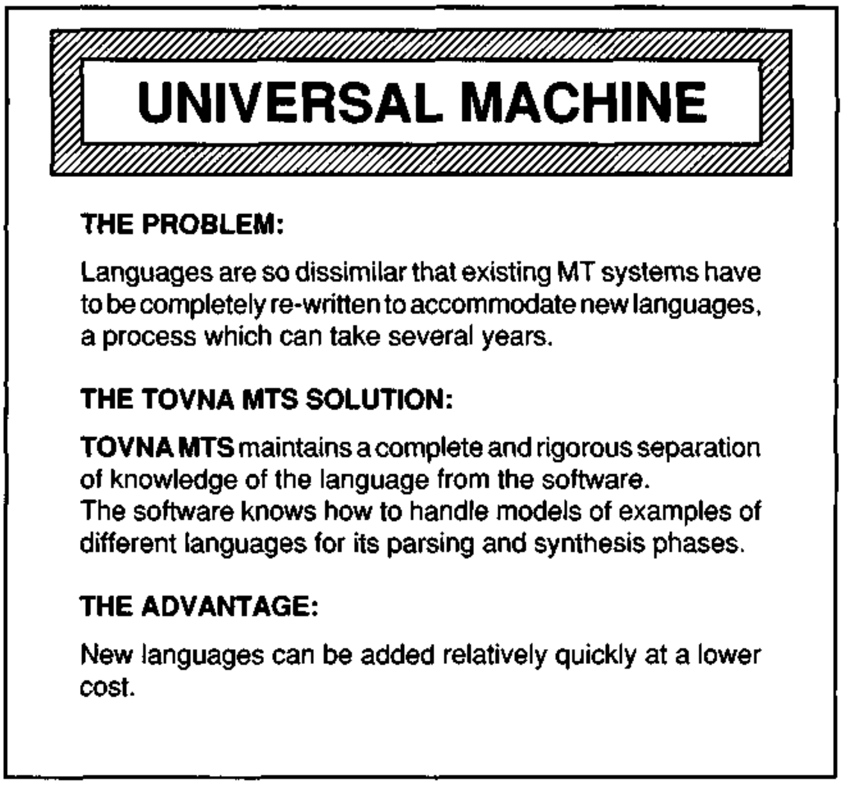
THE TOVNA MTS SOLUTION:

Sophisticated and easy-to-use menu-based screens enable the user to enter the required data accurately and quickly in a user-friendly environment.

THE ADVANTAGE:

The system is always up to date with the latest information, based on current user needs.

TOVNA MTS makes it easy to add words and phrases to dictionaries by providing direct access to the lexicon and dictionaries from the post-editor and/or any translation phase data bases with sophisticated and easy-to-use menu-based screens. The user enters data accurately and quickly in a user-friendly environment.



UNIVERSAL MACHINE

THE PROBLEM:

Languages are so dissimilar that existing MT systems have to be completely re-written to accommodate new languages, a process which can take several years.

THE TOVNA MTS SOLUTION:

TOVNA MTS maintains a complete and rigorous separation of knowledge of the language from the software.
The software knows how to handle models of examples of different languages for its parsing and synthesis phases.

THE ADVANTAGE:

New languages can be added relatively quickly at a lower cost.

The problem of adding new languages to an MT system is especially vexing because languages are so different from one another. Most existing systems have to be completely rewritten to accommodate a new language, a process which takes several years. Often, the new system has different capabilities and a different user interface, which can be confusing for users.

One of the most important features of TOVNA MTS is the rigorous separation between the software and the knowledge of a language. All the data describing languages is kept separate from the programs in the linguistic data base.

Because TOVNA MTS maintains a complete and rigorous separation of knowledge of the language from the software, new languages can be added relatively quickly.

The language's complexity is reflected not in the algorithms but rather in the rules and in the example-based language model. More complex languages simply have more rules and more examples in their models. The software is the same for all languages, and the system's capabilities and user interface are consistently maintained across all languages.

COMPATIBILITY

In addition to being language-independent (that is, the same software works with all languages), TOVNA MTS is also operating system-independent and machine-independent. In other words, TOVNA MTS can work with most commonly available operating systems and most commonly available computers. The fundamental operating environment requirements are:

1. a 32-bit processor (i.e., that the C compiler uses 32-bit integers);
2. 16 MB real memory and a virtual memory space of 24 MB;
3. 300 MB of disk space (for each language pair).

TOVNA MTS has been written in the ‘C’ programming language which runs on a wide range of computers and their operating systems. Choosing TOVNA MTS does not necessitate buying a new computer system.

THE TRANSLATION PROCESS

The TOVNA MTS translation process is divided into the following phases:

1. **Typography** divides the text into words, phrases, sentences and paragraphs and determines typographical attributes.

2. **Morphology** consults the syntactic dictionary lexicon and morphological rules to determine the morphological attributes of words and idiomatic expressions.

3. **Parsing** analyses sentences to determine the roles of words and phrases in each sentence.

4. **Transfer** substitutes target language words and phrases for source language words and phrases.

5. **Synthesis** constructs meaningful and grammatically correct sentences in the target language.

6. **Post-editing** enables the user to post-edit the translated document and to teach/improve the system using TOVNA MTS’s sophisticated bilingual text editor.

AN EXAMPLE OF THE TRANSLATION PROCESS

Consider the following sentence:

He saw the personal com-
puter manual on the table.

1. **Typography** collects all the typographic alternatives.

   computer or com - puter?

   Is it just one word that has been hyphenated because it appears at the end of the line, or is it a phrase which would be hyphenated in any case?

2. **Morphology** determines the correct typographical alternative and collects all the morphological alternatives by consulting the lexicon and morphology rules to determine the morphological attributes of words and phrases.

   saw (verb) or saw (noun) ?

3. **Parsing** in the first phase (syntactic) rejects most morphological alternatives and collects all the possible alternatives, while the second phase (semantic) rejects alternatives based on semantics and pattern statistics.

   He saw (the ((personal computer) manual)) on (the table.)
   or
   He saw (the (personal (computer manual))) on (the table.)

   Does the adjective 'personal' modify the word "manual" or the word 'computer'?

4. **Transfer** determines the correct alternative translation for each word/phrase based on context.

   table (furniture) or table (chart) ?

5. **Synthesis** constructs a grammatically correct and meaningful sentence from the target language elements produced by the transfer phase.

   Il a vu le manuel de 1'ordinateur individuel sur la table.

SOFTWARE PROGRAMS

Translating programs

The **TOVNA MTS** software is divided into modules which broadly correspond to the translation phases listed above. Each of the phases is accomplished by an individual program or set of programs. These programs are known as the 'translating programs'.

Building programs

Also corresponding to the translation phases is a set of 'building' programs which construct the linguistic data base for each language and language pair. For example, there is a set of typography building programs which builds the linguistic data base used by the typography translating programs. The translating programs do not include the knowledge of any of the languages they process. They only know how to compare text in a given language to the linguistic data base for that language. In this way, a complete and totally rigorous separation of software and knowledge is achieved.

LINGUISTIC DATA BASE

The linguistic data base consists of:

1. **A syntactic dictionary (lexicon) for each language.**

   The syntactic dictionary, also known as a lexicon, *lists the grammatical attributes* of each word or idiom in the language.

2. **A bilingual dictionary for each language pair.**

   The bilingual dictionary gives the *context-based translation* for each entry (word or idiom).

3. **A bilingual 'phrase table' (synthesis model) of examples for each language pair.**

   The bilingual 'phrase table' gives *the translation rule* for each phrase pattern in the source language.

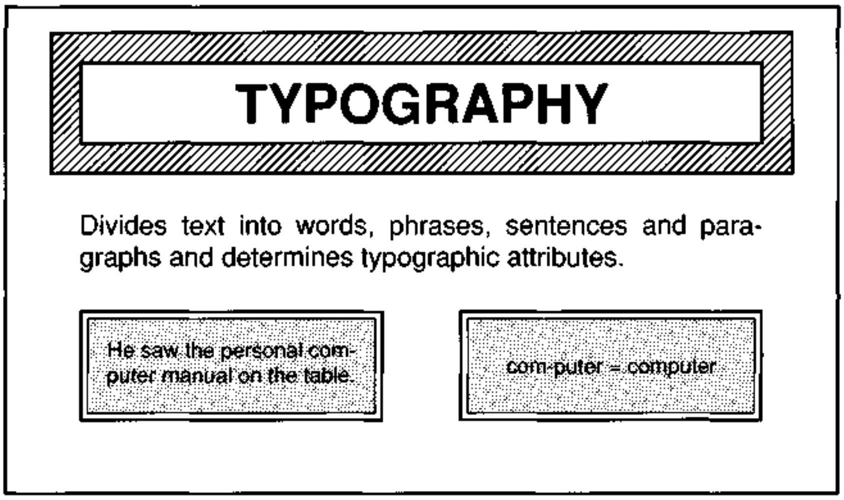
4. **A parse model for each language.**

   The parse model is used in analysing the possible syntactic structures of the source text.

5. Miscellaneous files for each language.

These files describe typographical and morphological rules, lists of attributes and roles, and automata.

TYPOGRAPHICAL ANALYSIS



TYPOGRAPHY

Divides text into words, phrases, sentences and paragraphs and determines typographic attributes.

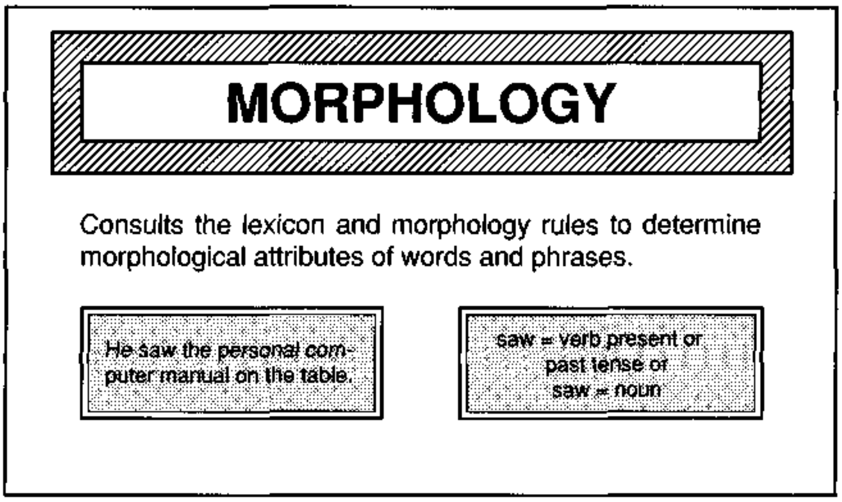
He saw the personal computer manual on the table.

com-puter = computer

The typographical analysis phase breaks the source text into words, phrases, and punctuation based on the typographical rules specified for the language. In the case of ambiguous text, typographical analysis identifies all the alternatives and passes them on to the next phase, morphological analysis. The typographical attributes of each word are added to the text. Typesetting and control codes are identified as such and are saved in order to be inserted into the target text in the appropriate places so that it can be directly typeset without further editing.

An example of a typographical ambiguity might be a hyphenated word at the end of a line, for example, pre-editing. It may be that the word is ‘preediting’ or perhaps ‘pre-editing’, that is, the word would be hyphenated even it is were in the middle of a line.

MORPHOLOGICAL ANALYSIS



MORPHOLOGY

Consults the lexicon and morphology rules to determine morphological attributes of words and phrases.

He saw the personal computer manual on the table.

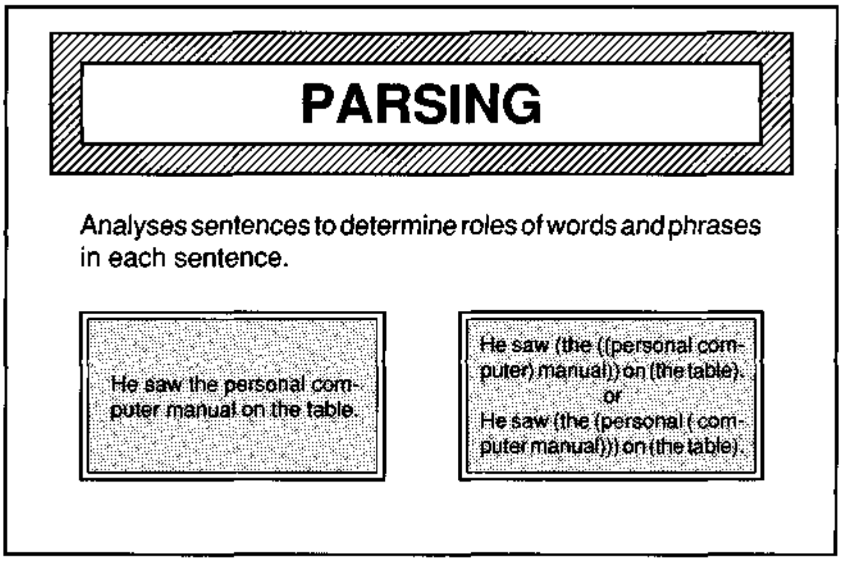
saw = verb present or
past tense or
saw = noun

Based on the morphological rules of the language and the entries in the syntactic dictionary, morphological analysis determines which of the alternatives developed by the typographical analysis phase are correct. The morphological attributes of each word are added to the text.

An example of a morphological ambiguity is a word such as 'table', which might be a verb or a noun, depending on context.

Morphology analyses each of the words in the text and determines what lexical entries they represent by applying the morphological rules. For example, one rule specifies that an English word ending in 'ed' may be the past participle form of a verb. Another rule specifies that the prefix 're' is pre-pended to a verb (to signify repetition). Morphology discovers that the word 'red' is not the past participle of the verb 'r' because there is no lexical entry 'r'. Nor is 'red' a form of the verb 'd' because there is no lexical entry 'd' in the syntactic dictionary. So morphology correctly identifies 'red' as an adjective, since 'red' appears in the syntactic dictionary as an adjective.

PARSING ANALYSIS



PARSING

Analyses sentences to determine roles of words and phrases in each sentence.

He saw the personal computer manual on the table.

He saw (the ((personal computer) manual)) on (the table).
or
He saw (the (personal (computer manual))) on (the table).

The purposes of parsing are:

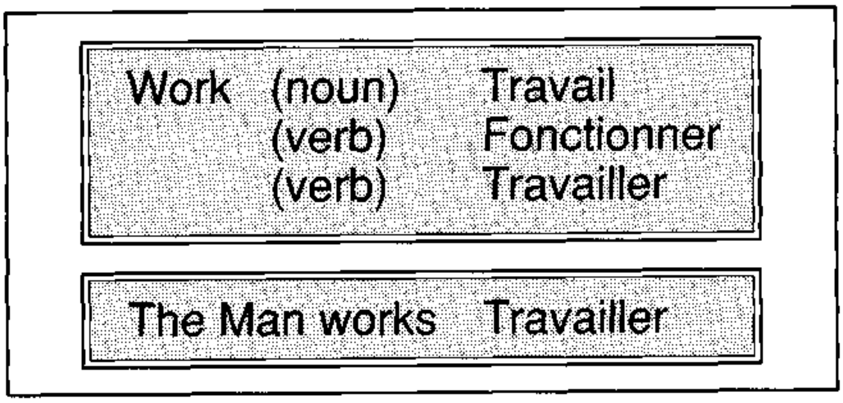
a. To resolve morphological ambiguity.
b. To resolve ambiguity in the translation of a word.
c. To determine sentence structure.

Resolving morphological ambiguity

Consider the phrases ‘they work’, where ‘work’ is a verb and the phrase ‘the work is hard’, where ‘work’ is a noun. It is the job of the parser to determine the part of speech of ‘work’ and thus to resolve the morphological ambiguity.

Resolving ambiguity in translation

The verb ‘work’ might be translated into French either as ‘travailler’ or ‘fonctionner’, depending on whether the agent of ‘work’ is animate (‘the man works’ – translated ‘travaille’) or inanimate (‘the machine works’ – translated ‘fonctionne’).



Work	(noun)	Travail
	(verb)	Fonctionner
	(verb)	Travailler

The Man works	Travailler

It is the parser which identifies the agent of the verb.

Determining a sentence structure

In order to construct a correct sentence in the target language, *one* of the following is necessary:

a. Extract from the source sentence a formal structure (logical form) and construct the target sentence according to the target language rules, the ‘interlingua’ approach. This approach uses an artificial middle language. All target and source languages are translated into and out of this middle language. In this case the parser constructs a formal structure from the surface structure.
b. Use transfer rules to get from a source language to a target language. Parsing is necessary with this approach (the transfer approach) because the number of possible different sentence structures is too large to enable the system to define a transfer rule for each one. TOVNA MTS therefore parses the sentence into its constituent phrases and applies the transfer rules to the phrases. This approach is practical because the number of possible different phrases is much smaller than the number of possible different sentences.

TOVNA MTS uses the transfer approach.

In summary, the functions of the TOVNA MTS parser are:

a. To break up a sentence into its constituent phrases; and
b. to determine, for each phrase, the thematic relations between its elements.

PROBLEMS IN PARSING

There are two fundamental problems in the development of a parser:

a. completeness;
b. ambiguity.

Completeness

A parser must recognise all the possible sentence structures in a language. Each language has an enormous number of possible surface sentence structures. The grammars of most parsers describe only a fraction of a language's possible sentence structures and most texts (especially technical texts) contain sentences which most parsers do not recognise.

Ambiguity

The analysis of a sentence's structure cannot be accomplished with syntactical rules alone. Consider the sentence:

Fruit flies like bananas.

Without a semantic rule stipulating that 'fruit' is unlikely to be the agent of 'flies' there is no way to determine the correct analysis of the sentence.

Completeness resolution

TOVNA believes that all the possible sentence structures in a language cannot be described by a reasonable sized system of rules. An algorithm has therefore been developed which automatically constructs parsing rules from examples.

In this way, a parsing model is constructed for each language *in an incremental fashion.* Tovna linguists parse a number of sentences 'by hand', and TOVNA MTS 'learns' the structures which make up these sentences. As the learning process progresses, the linguist need not completely analyse new sentences, but rather, with the aid of an interactive program, can follow the parser as it attempts to analyse new sentences, intervening only when the parser is unable to parse a sentence successfully. Every new parsing that the linguist authorises updates the model and adds to the parser's ability. Since the number of possible phrase structures is not very large, this process converges relatively quickly.

The prodigious advantage of this method (learning by example) is that the enrichment of the model at each phase is, from the linguist's point of view, a simple process. To add a new structure the linguist need only analyse a phrase of that structure. This approach is in marked contrast to the rules-based approach where a linguist who wishes to add a new rule must analyse all the possible interactions between the rule to be added and existing rules.

It is only natural for such other systems to reach rather quickly the saturation point (where no more rules can be safely added) and thus never approach the goal of completeness.

Another important advantage of the TOVNA MTS method of building a model from examples is that building a model for a new language requires no new software. One simply starts, as described above, with the agreement rules for the new language and builds upon them using the same programs used for all other languages. In other words, Tovna linguists teach the machine a new language the same way a child learns a language.

Ambiguity resolution

As mentioned earlier, there are often several syntactical analyses of a sentence. Of these, only a fraction are semantically possible and correct. The goal therefore, is to select from among these the one which is semantically 'most reasonable'.

The great difficulty in solving this problem is the enormous amount of semantic knowledge needed to determine if an analysis is semantically correct. This semantic knowledge consists of the permissible thematic relationships between words; for example, the fact that the relationship subject-verb is an unlikely one for the words 'fruit fly' but a most likely one for the words 'airplanes fly'. Note that the issue is further complicated by the fact that not all correct relationships are equally correct; sometimes the system must be able to select the best or most reasonable from a group of correct relationships.

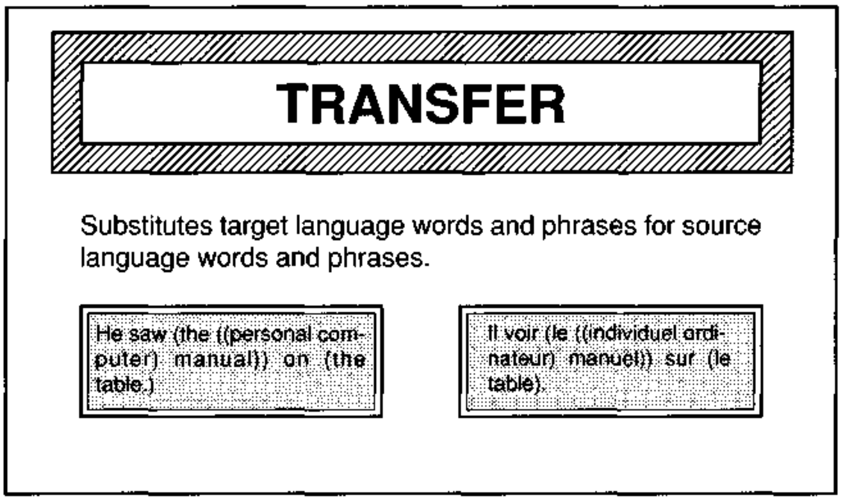
In the solution of ambiguity resolution, as with completeness, TOVNA MTS uses the approach of learning from examples. TOVNA MTS 'learns' from each analysis of text which thematic relationships are common and uncommon. Thus, if many texts about a particular subject are analysed, TOVNA MTS accumulates a great deal of semantic information about words frequently used in texts dealing with that subject.

TOVNA MTS also defines semantic categories. Any word in the lexicon (syntactic dictionary) can be described as belonging to a semantic category, enabling TOVNA MTS to determine the correctness of thematic relationships between words for which no such relationship has been defined.

For example, if TOVNA MTS knows that the relationship subject-verb for the words 'water flows' is correct, and both 'water' and 'soup' are in the category of 'liquid', then the relationship is correct for the words 'soup flows' as well, even though it has never been explicitly defined as such.

Up to this point, only the source language is relevant. That is, no matter what the target language is, the typographical and morphological analyses as well as the parsing are the same. It is only from the transfer phase on that the target language becomes relevant. For example, if the same document must be translated into several languages, the first three phases of the translation need be performed only once, achieving a considerable time saving.

TRANSFER



TRANSFER

Substitutes target language words and phrases for source language words and phrases.

He saw (the ((personal computer) manual)) on (the table.)

Il voir (le ((individuel ordinateur) manuel)) sur (le table).

The purpose of the transfer phase is to find, for each word (or expression) in the source language, the corresponding word (or expression) in the target language with the help of a bilingual dictionary. The most important difficulty is that the bilingual dictionary contains, for each source word, more than one translation, and TOVNA MTS must determine the correct one based on the subject or the context of the word.

There are several distinct problems here. The first is that of homographs. Homographs are two (or more) different words which have the same form, for example ‘board’ meaning a circuit board in a computer and ‘board’ meaning a board of directors. (It is perfectly reasonable to view these as two distinct and unrelated words which happen to be written identically). In this case the correct identification of the part of speech must be made *before* the transfer phase. This identification is part of the process of resolving the ambiguity of the sentence in which the word appears and is described in the previous section.

A second problem arises when the target language uses different terms for the same word in the source language. For example, in French the words ‘travailler’ and ‘fonctionner’ are both used in place of the English verb ‘work’ depending on whether the agent of the verb is animate or inanimate. (In this case, the bilingual dictionary specifies conditions when one translation is to be used in preference to others).

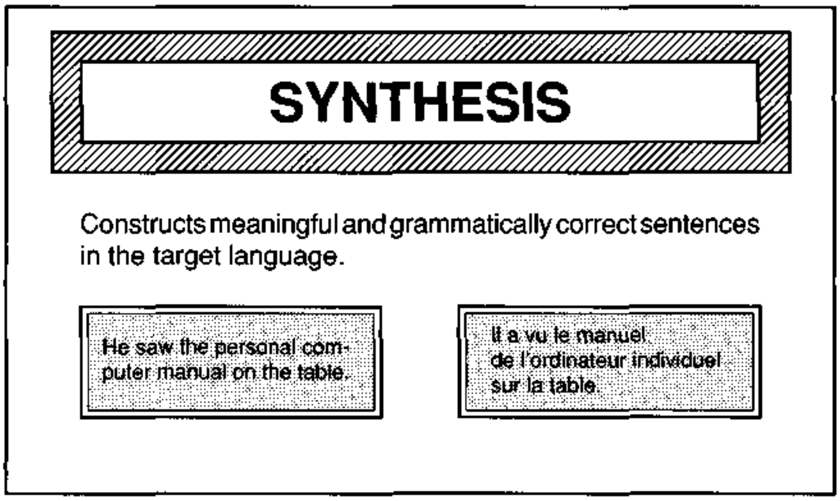
Translating prepositions is another common problem. The **TOVNA MTS** solution is to include in the bilingual dictionary translations of not only verbs but also of the different valid verb-preposition combinations, so that the appropriate preposition will be chosen for the translation. Words derived from verbs are handled in the same way.

The problem of translating a word such as ‘run’ is solved in a similar manner. The bilingual dictionary has separate entries for ‘run <object>’ and ‘run <no object>’.

In addition, the bilingual dictionary contains idioms of arbitrary length.

BILINGUAL PHRASE TABLE OF EXAMPLES

(Synthesis model)



SYNTHESIS

Constructs meaningful and grammatically correct sentences in the target language.

He saw the personal computer manual on the table.

Il a vu le manuel de l'ordinateur individuel sur la table.

The synthesis phase constructs meaningful and grammatically correct sentences from the target language by rearranging words, determining morphological attributes of words, and adding or deleting words according to the synthesis rules.

In addition, a set of synthesis patterns (the bilingual phrase table) is built using an interactive program. Here too, TOVNA MTS learns by example. It was concluded that constructing precise rules for a target language structure based on the source language ‘analysis tree’ is too difficult a task for the user because of the tremendous variety of possible source language structures, as well as the great differences between source and target language structures. For this reason, TOVNA MTS here too, takes the approach of few rules and many examples. In this way the user can easily enhance the phrase table model by simply adding more examples. The only rules needed are the agreement rules of the target language. On this foundation, the user specifies examples of synthesising a target phrase from a source phrase.

During post-editing, the source language and target language versions of the same text are displayed, allowing the user to match corresponding sentences and phrases. Alternatively, the bilingual phrase table can be directly edited. The user may specify the correct translation of a phrase, which includes the rearrangement of word order, and the deletion and insertion of words.

EXAMPLES

(1) changed word order

‘personal computer’ → ‘ordinateur individuel’

(2) added word(s)

‘computer manual’ → ‘manuel de 1’ordinateur’

(3) deleted word(s)

‘they continue on their way’ → ‘ils continuent leur chemin’

In practice there are frequently many target structures appropriate to a given source structure. When to choose one rather than another depends on the attributes of the source phrase and sometimes on the context in which the phrase appears (that is, context outside the phrase).

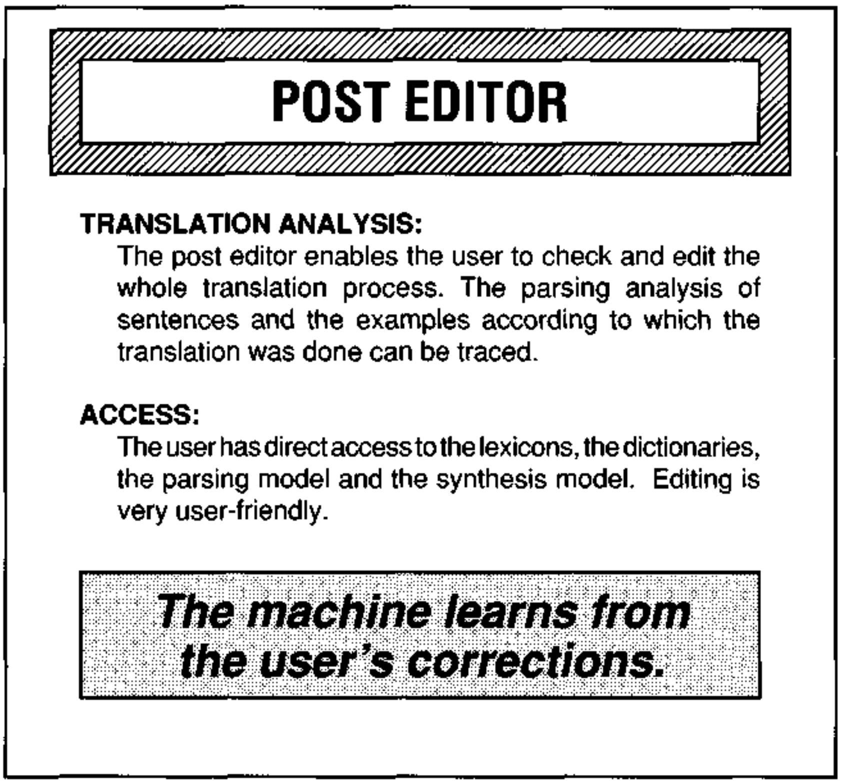
For example, the structure adjective-noun can be translated into French as adjective-noun or noun-adjective, depending on whether the French adjective is one which comes before or after the noun. The French lexicon (syntactic dictionary) specifies, for each adjective, whether it comes before or after the noun. The bilingual phrase dictionary contains the information that for an adjective-noun phrase, the word order of the target phrase depends on the attributes of the adjective as specified in the lexicon.

Thus, the number of examples is rather large, and each example contains detailed conditions specifying when it is to be used.

The user identifies incorrectly translated phrases (where the error is in the structure of the target phrase) and can give a new example to the table model. TOVNA MTS extracts the correct synthesis rule from the example and adds it to the model, enriching the model and improving the quality of subsequent translation. The user can influence the 'style' of the translation in this way.

Once the target phrase has been synthesised on the basis of the synthesis model, there remains only the last step in the synthesis process: morphological synthesis, which is guided by the morphological attributes determined for each word in the earlier stages. Morphological synthesis is the opposite of morphological analysis. For example, the morphology rules specify that to determine the root form of an English noun, the suffix 's' should be deleted. This same rule, applied in reverse, specifies how to form the plural of an English noun (add an 's').

POST EDITING



POST EDITOR

TRANSLATION ANALYSIS:
The post editor enables the user to check and edit the whole translation process. The parsing analysis of sentences and the examples according to which the translation was done can be traced.

ACCESS:
The user has direct access to the lexicons, the dictionaries, the parsing model and the synthesis model. Editing is very user-friendly.

The machine learns from the user's corrections.

The main functions of the post editor are:

1. tracing the system's logic in the translation process;
2. teaching;
3. post-editing.

Teaching

The TOVNA MTS system is a learning system. The teaching is done through the post editor. The system selects the correct translation by choosing one of the examples found in the parsing and synthesis models. In the case of an incorrect translation, the user can correct the translation of an incorrect phrase and then add that correct phrase to the model of examples maintained by the system. All of this is done through the post editor. From then on the system will know how to translate new sentences which are similar in structure to this new example.

Tracing the systems logic in the translation process

The TOVNA post-editing program, called ‘post-edit’, displays both the source and target texts side by side on the screen in adjacent vertical windows. The source text is on the left and the target text is on the right, so that each source sentence is displayed next to the target sentence. The user can freely navigate in both texts, synchronise the display, or scroll texts individually, find a specific word in either text or selectively replace words.

For example, the user can ask TOVNA post-edit to find the source word or phrase which corresponds to a specific target word or phrase or vice versa. The system allows the user to see all the phases of translation and have the system explain the results of each phase.

The user has access to all phases of the translation process and can make the following changes:

1. Select an alternative morphological analysis.
2. Select an alternative parsing for a phrase or sentence.
3. Add or change examples in the bilingual phrase table.
4. Update the monolingual syntactic dictionary (lexicon).
5. Update the bilingual dictionary.

Due to the teaching process, *the system's accuracy constantly improves.* The system will require less and less editing and teaching as it becomes more and more accurate with use.

Translating and editing

TOVNA MTS is very user-friendly. Commands and instructions are universal and readily recognisable to translators and editors. While the format and logic remain the same, these commands and instructions can appear in any language in the database. With TOVNA MTS, there is an optimal Man-Machine Interface which achieves impressive results in a very short time. Translators can work either online (interactively) or in the automatic batch processing mode (overnight translation of large documents) followed by interactive post-editing.

TOVNA MTS *can increase output, reduce overheads and accelerate turnaround time.* Editors increase their output because TOVNA MTS ensures that documents are complete, consistently-written and spelled properly. It automates many repetitive processes and lets translators spend more of their time on language and style instead of retranslating similar, identical or repetitive texts. TOVNA MTS enhances productivity while ensuring the highest quality translation possible.

ADDITIONAL STRENGTHS OF THE TOVNA MTS SOLUTION

1. The system can create separate "translations" according to subject matter by defining the possible subject fields relating to specific words.
2. There is no limit to the number of words and/or idioms which can be stored in monolingual lexicons and bilingual dictionaries.
3. There is no limit to the number of possible translations which can be assigned to a word. Even the same part of speech can be assigned more than one translation (in case of a semantic difference).
4. There are no limitations on the entry of words into the lexicons and dictionaries.
5. TOVNA MTS dictionaries can be transferred or merged from one site to another.
6. The system can be interfaced with all popular word processing software.
7. TOVNA MTS output can be sent to typesetters with all necessary instructions encoded. The system ensures that typesetting instructions in the original input will automatically be retained in the translated version.

USER BENEFITS

Translation involves time-consuming and repetitive manual tasks that are routinely but necessarily performed by translators and editors.

TOVNA MTS is designed to improve dramatically the productivity of translators and editors by substantially reducing pre-editing and post-editing translation time and cost, while ensuring consistent, accurate, and up-to-date translations.

Cost/effectiveness analyses vary with each customer's labour costs, translation volume and productivity. TOVNA MTS should save customers money by reducing per word translation costs between 25 per cent to 75 per cent once the system is fully trained by a user, approximately a six-month process. In addition, TOVNA MTS should permit a translation organisation to increase the volume of translation substantially without increasing manpower.

Within the translation market, translation costs vary by language pair and sector. When a translation service bureau performs the work, the price of a technical manual can reach as high as $62.50 (US) or more per page, calculated on the basis of $0.15 - $0.22 per word for English to/from French, Spanish and German. This is the standard price range in the translation business. Other European languages (Dutch, Portuguese, Russian and Italian from English) are more costly: $0.18-$0.25perword. 'Exotic' languages (Arabic, Chinese, Japanese and Korean from or to English) cost $0.25 - $0.35 per word.

In contrast, the internal costs of translation, when they include the allocation of indirect overhead costs, are estimated by translation departments of Canadian companies to range from $0.12 to $0.25 per word from English to/from French. The labour cost of a technical translator or editor can amount to $40,000 to $60,000 per year depending on his level of technical background and years of experience. In the translation business, the volume of translations performed by a translator is estimated to range between 1,000 - 3,000 pages (250,000 - 750,000 words, assuming 250 words per page) per year.

An editor typically performs quality review for three to five translators. An editor is a highly experienced translator with greater technical skills. On the average, an editor reviews and revises 30 - 50 pages per day, polishing the style, correcting errors and checking for completeness and consistency in the translation.

An annual volume of translation of 7,000 or more pages per year appears to justify the use of TOVNA MTS by an in-house translation department employing two translators and one editor. The estimated annual cost for such an operation excluding the purchase cost of the software is estimated at less than $ 170,000 for two translators (users) and one terminologist (super user) who would also function as an editor.

If used according to the recommended specifications, TOVNA MTS should enable the per word cost to fall below 10 cents per word when the volume exceeds 1,750,000 words per year. At a volume of 2 million words per year this 'translates' into an annual saving of $100,000. Based on these assumptions, increased translation volume and/or productivity would lead to additional savings.

The payback period for a TOVNA MTS licence, when properly utilised, is approximately 18 months at a volume of 2 million words translated per year.

FUNCTIONAL SPECIFICATION

FOR
A TWO-USER ENGLISH TO FRENCH
VERSION OF TOVNA MTS

1. **ESTIMATED PROCESSING TIME.** The estimated processing times below assume implementation on a Sun Microsystems 3-60 with 16 MB for RAM.

A. **Batch Mode.** The number of words processed per hour in MTS Batch Mode depends on the number of users working on the system simultaneously. For one text an average number of words processed per hour is estimated at around 3600 (assuming that a 10 word sentence is processed in 10 seconds).

When a second text is processed simultaneously, the system should process both texts in less than double the combined processing time. Such batch processing can also be conducted with interactive work, though response time, as described below in Paragraph B, will be slower.

B. **Interactive Mode.**

1. The average time required to make a lexical entry in one of the lexicons (English or French) for a trained user is 1 minute or less.
2. The average time required to update the dictionary depends on whether the source and target words have already been entered into the respective lexicons. If the words are already in both lexicons the average time for a trained user is 1 minute or less. If such lexical entries are missing then an additional 1 minute or less (as noted in B.1) is required.

2. FUNCTIONS

A. Post Editor Mode

1. The system allows direct access from this mode into the phrase dictionary, the bilingual dictionary, and into the source and target lexicons. The system allows each of these functions to be updated without returning to the main menu.
2. The user can review the parsing analysis of the source text. In addition, a trained user may ask the system to display an alternative parsing analysis if it is available.
3. The split screen allows the system to display equivalent words and phrases in the source and target languages.
4. The system allows both synchronised screens and full page review of either source or target texts.
5. The system includes a word-count function for either source or target texts.

B. Phrase Dictionary

1. The system allows direct access from this mode into the bilingual dictionary and the source and target lexicons. The system allows for the updating of these functions without returning to the main menu.
2. The system allows multiple target entries for each source phrase.

C. Bi-Lingual Dictionary

1. The system allows direct access from this mode into the source and target lexicons. The system allows for the updating of these functions without returning to the main menu.
2. The system allows multiple target entries for each source entry (even for the same source lexical entry).

AUTHOR

Ami Segal, Tovna MTS, Beit Ha-Arave St 28, Jerusalem 93385, Israel.