INTERNATIONAL JOURNAL OF TRANSLATION
Vol. 15, No. 1, Jan.-June 2003

Inside a High Quality English-Spanish MT Engine

RICARDO ARGÜELLO MÁNTICA
Word Magic Software Corporation

INTRODUCTION

One of Arthur Bloch's aphorisms on computers says:

> *A computer will do what you tell it to do. It won't do what you want it to do.*

When one mentions machine translation at least two strong reactions are usually expected from an educated audience. One of them is a keen, instantaneous interest and curiosity on the subject, similar to the thrill of traveling through unknown lands and coming face to face with the seductive magic of some kind of 'machine intelligence', which - incidentally - we think does not exist at all and is still light-years away, based on our present-day technology. This so-called intelligence is nothing but a sophisticated collection of rules, piles of information and a web of algorithms combined with dazzling speed for manoeuvering ones and zeroes.

The second probable reaction is one of flat skepticism. Past experience with machine translation has generally been deceptive, to say the least. Quite comprehensibly, this feeling is particularly strong among translators, who probably feel either deceived or threatened by the new intruder, quite a human reaction.

On the other hand, the machine translation industry players are a passionate group with a missionary zeal. Curiously enough, we have not given up nor run away in complete despair at the sight of such harrowing results from our so-called translation systems. We are even loath to accept that MT, even after more than 50 years from its inception, is still in a ludicrous position compared to other branches of computer science.

If airplanes or automobiles were manufactured with such low accuracy standards as is typical in MT, there would exist no automotive nor aeronautics industries at all. And true enough, it can be said that a Machine Translation industry as such still does not exist in the world today. There are not even any accepted standards for correct performance, nor proven methods for numerically measuring such performance.

At this stage, other probability-based systems such as OCR, TTS and Voice Recognition are far more accurate than language translation.

Millions of dollars have been spent in MT software research, mathematical modeling and linguistics analyses. There were some early applications developed back in the 1960s, but the results were so poor that the Government stopped all grants destined for MT research in the US. Noam Chomsky finally anathematized machine translation, labeling it as 'unfeasible'. Perhaps an accurate appraisal at the time... and perhaps an accurate appraisal at present, according to the following conclusions reached by Cloutier & Associates after a study of the market for computer-assisted translation:

1) *MT is not considered a corporate priority!... there are no industry case studies that demonstrate a quantifiable ROI.*
2) *The market is fragmented with multiple needs and requirements... different language pairs, high quality versus gist, general purpose versus domain-specific.*
3) *MT has been over-hyped in the past, making most potential end-users calloused of current claims. Education is needed to show businesses how MT can really help them today and to reset expectations.*
4) *The* Web *has become a double-edged sword by creating more cross language communication, creating more need for translation and raising awareness of MT, but the many bad, free on-line MT engines have left most people feeling that MT is still not ready for prime time.*

But all is not bleak in this awesome scenario. If we reverse the logic and achieve a truly accurate electronic translator, we may assume that a Machine Translation Industry will immediately emerge with the force of a tidal wave. Possible applications are almost innumerable and the demand is there, waiting for a practicable system.

At present, there is no market for MT because there is no product.

> *Given the high degree of creativity and complexity involved...unfortunately there is no "black box" system that can remove all the difficulties inherent in language translation."* CALICO Review on ESI, Universidad La Coruña, Spain 1999

Let's consider a few reasons why MT is in such a state of underdevelopment.

I. WHAT ARE THE MAIN DIFFICULTIES IN MT?

"A linguist is anyone who can make himself misunderstood in more than one language." Arthur Bloch.

It would seem at first glance that interpretation of a spoken or written language, should not be a greater burden on a computer than optical character recognition (OCR), text to speech processing (TTS), or even voice recognition (VR). However, if we stop and think about it, we will see that the number of possible permutations involved in solving even a simple sentence is astronomically higher than for the other tasks mentioned.

In OCR we deal with only 52 lower and upper-case characters plus a few diacritical signs. In we process sounds formed only by combining some 12 vowels sounds with 21 consonants. Allowing for as many variations and permutations as you will, the total number is still within reasonable limits. Both OCR and VR have evolved to levels of 90%+ accuracies.

However, considering that in the English language there are close to a million words, if we discard restrictive combinational rules the possible number of permutations can rocket way past the capacity of any imaginable computational system. Even in a relatively short sentence containing 15 words, the number would be something like 10 raised to the 75th power, which is a larger figure than the estimated number of atoms in the entire universe.

You may argue than nobody would expect to encounter all sentences originating from an entire one-million word dictionary. We would most likely restrict our dictionary to the 10,000 most commonly used words. Even so, a sentence formed with only 4 words would involve about 10 raised to the 16th power *of liberal* permutations. Just reading through 10^{16} and allowing one second for each sentence would take a man 318 million years without taking a single break. It would take a typical computer of, say 1 GHZ, about 50 thousand years to process all those permutations. Just add one more word in your sentence and the time required by the computer will soar to more than 300 million years

This is why machine translation cannot be done by brute statistical methods alone. Intelligent restrictive rules have to be put in place to check and hold back unrestrained association of words and thus limit the possible permutations.

The statistical approach was unsuccessfully tried at the onset of machine translation by several large companies, and it is now coming back with renewed force through a system called EliMT. This new system claims that feeding the machine with all novels, papers, documents and news

media available through the Internet in a worldwide coordinated effort, and also feeding it their respective translations in order to create a gigantic corpus, one might expect to obtain a kind of Translation Memory database capable of translating all language pairs at 'at least a Systran level' of gist understanding.

In our opinion, there is definitely not much hope for this kind of research, particularly if we consider languages such as English which is highly dependent on semantics, idiomatic expressions and word clusters, or worse still, Spanish which is sensitive not only to semantics and idioms but also to what we call *idiomatic constructions*, and *pragmatic dependency*. To further complicate matters, Spanish is also sensitive to regional differences and its verbs can mean totally different things depending upon the position and number of the *clitic pronouns.*

Can something so complicated become even more complicated? It certainly can!

There is a general tendency in most languages to become more and more lax as they evolve and embrace the cultural elements of the population. There is nothing wrong with a language embracing elements like idioms, colloquialisms, technical jargon, or even street language. There is nothing wrong even with inventing new ways of assembling phrases, which is precisely what we call *idiomatic constructions,* expressions, for instance: *'what a nice car!', 'a two-footer', 'doctor-recommended', 'strong-chested'* or *'my son is ten now'* ...All of these non-standard, non-grammatical and folkloric ways of expressing things actually enrich languages, make them more versatile and give them more power to communicate with every strata of the population. But they make it much more complicated for the computer. Now we have: scientific language for the scientist, metaphoric expressions for the literary person and the poet, telegraphic, quasi-symbolic language for the busy executive, and slang and street language for the man on the street, just to name a few language types. What else can we ask for?

But there are many other complications that have to be dealt with even before the sentence enters the grammatical parser.

There is a current tendency to oversimplify sentences in business and informal communication, many times leaving out vital elements and thus ending up with a lame syntactic construction. This is typical now of communications through email.

Therefore, the sentence cannot be fed into a parser as such. It must be clarified grammatically and syntactically first through a series of preparsing modules capable of identifying all current idiomatic constructions and

replacing them with straightforward orthodox sentences. When you feed a translator a sentence such as:

Hope she is well.

the typical outcome in Spanish in all popular MT applications is something like

Esperanza que ella está bien.

A more advanced system has to be capable of identifying this unorthodox elision and adding the correct pronoun that was left out by the writer. The pronoun could be either 'I',' you ' or 'we'.

i. *I hope she is well.*
ii. You, hope she is well. (Imperative)
iii. Let's hope she is well.

Our system will choose option (i), "*I hope she is well - Espero que ella esté bien"*. We will come to these cases later on in the article.

In Spanish the problem is worse because leaving out accents and other diacritical marks, as is customarily done in emails, implies a total change of meaning and/or part of speech. Therefore, our system contemplates another module we call *ambiguity checker,* which indicates to the User the Spanish words which can be written with or without accent, a condition a normal spell checker cannot identify because those ambiguous words are correct in both forms.

Take for instance the term *comete,* in Spanish. Thus written, without the accent, it is a conjugated form of the verb *cometer,* which means *to commit.* Whereas the same word with an accent, *cómete,* is a conjugation of a totally different verb, *comer,* which means *to eat.* So, what does the email mean when it says *'Comete eso'!*

In addition, there are thousands of possible ambiguities also in nouns. *Sabana,* and *sábana* mean two totally different things: *savanna* and *bedsheet* (bedding). Our ambiguity checker will highlight those words for the User indicating their respective meanings so that he or she can make a correct choice before inputting the text into the translator. Otherwise, the old formula will invariably apply:

GI = GO : Garbage In = Garbage Out

Nevertheless, there is more to the problem of ambiguity. About 60% of the total words in English are polysemous; that is, they have two or more parts of speech, and each part of speech can have two or more meanings,

depending on context and their respective position within the sentence. A noun could be a verb or an adjective, or an adverb, or an interjection, or even a preposition.

This condition makes grammatical parsing a very complex task. A parsing system has to include quite a large number of sophisticated rules in order to determine the correct part of speech of each ambiguous word within a sentence.

But the real difficulty only begins here. Once the correct part of speech has been determined for a particular word, the system has to be able to choose the correct meaning when confronted with a polysemous word. A great deal of artificial intelligence is needed for this kind of task, and currently a large number of researchers and linguistics groups around the world, among them several financed by Microsoft Corporation, are pursuing solutions to this seemingly insurmountable problem.

A normal translation program would never tackle this magnitude of complication, and Word Magic does not either. However, we are already in the process of implementing knowledge databases and developing our own intelligence system to try to disambiguate words of this type.

At this stage, Word Magic is able to disambiguate polysemous words only in certain limited cases, for instance, in the sentence:

I am teaching this bat to fly.

Word Magic will disambiguate bat and correctly choose *murciélago* (animal) instead of *baseball bat.* It will translate:

Le estoy enseñando a este murciélago a volar.

However, if you enter: "*I am playing baseball with this bat,"* it will logically choose *baseball bat* and translate accordingly.

At present, no commercially available translation program will even attempt to disambiguate nouns or verbs. In this sense, Word Magic is many years ahead of the rest of the available MT systems in the market.

II. WHAT IS WORD MAGIC'S APPROACH TO MT?

Word Magic Software's newest release, ESI PRO Version 4.0 (shown below), together with a set of translation tools, mirror-thesaurus, dictionary and conjugator TDT PRO Version 4.0 (Translation Dictionary & Tools), contains outstanding improvements in the translation engines in both directions, (Release date: March 31th or early April 2003.)

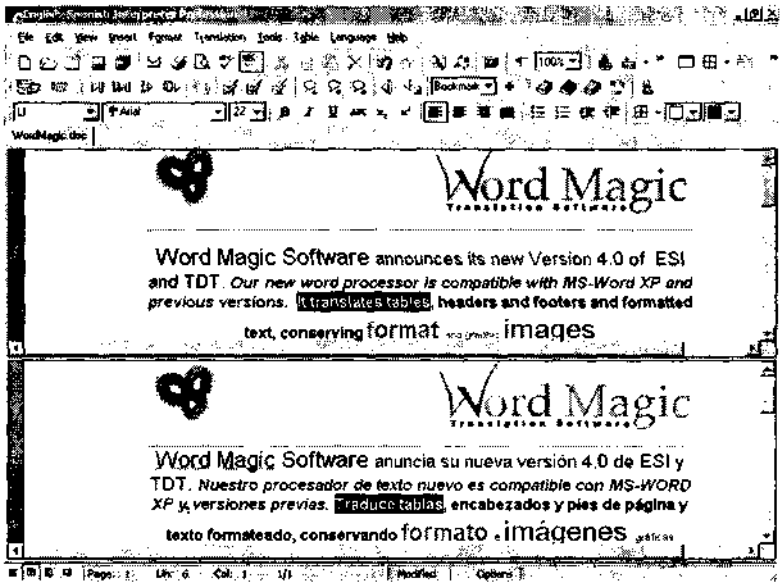

Fig. 1. Main Screen of ESI 4.0, showing automatic translation of formatted text, graphic images, highlights and colors.

Word Magic Software is a company that specializes in just two languages: English and Spanish. Thus, our investigative efforts can actually focus on the finer details and the hidden, almost esoteric, rules that apply to language interpretation and translation, which depend on semantics, pragmatics and context recognition. It would be impossible for any one company to mass-produce translation software for several language pairs with an acceptable degree of accuracy.

I am tempted again to borrow from Arthur Bloch's humorous quotes *"An expert is a person who has committed all the possible errors in a very limited field of study"*. So, in this sense we consider ourselves real experts in English to Spanish translation.

Our philosophy is that we should deal with only one language pair if we want to produce an application worthy of being called a *translator*, something that would serve not only to amuse people or give a gist or vague idea, or even worse, a distorted resemblance of the original document when translated into the other language, but rather as something that may be of real assistance to the human translator using it. We have produced such an application and it is called ESI, which is the acronym for English Spanish Interpreter. As such, we view ESI Version 4.0 not only as a MT system, but mainly as productivity tool.

The English Spanish Interpreter (ESI) is a text translator capable of translating whole documents from English to Spanish and vice versa at a click of a button in Automatic mode, or sentence by sentence in Interactive Mode where the User has complete command over the translation process.

ESI has been on the market since 1998, but its creation actually began in 1988. We use our own dictionaries, especially designed and constructed for ESI, containing not only the largest number of translation references of any other English-Spanish dictionary in computer form (850,000 entries), but also the largest number of electronically-inferred synonyms (2,000,000 in each language) and also the largest collection of idiomatic phrases, phrasal verbs, particle verbs, adverbial expressions and set-phrases, all with their bidirectional translated correspondences in the other language (225,000).

We have no intention of ever replacing the professional human translator. Our aim is only to produce a practical tool, one capable of rendering automatic translations of reasonably good quality as a first draft, which in turn can be easily edited in a semiautomatic, interactive mode guided by the human operator.

All of these considerations, plus the permutational impossibilities described earlier, led us to discard the statistical method from the very start and set about actually seeking *patterns and rules* capable not only of dealing with the syntax and grammar of the sentence, but also with its semantics, *category field* and *pragmatic content*, all concepts introduced for the first time ever by Word Magic Software, and which we consider to be an important breakthrough in the realm of Artificial Intelligence applied to machine translation. We will explain what we mean by those terms later on.

We have to confess that we did not start our research based upon any existing theory or model. We did everything from scratch, even the dictionary. We followed no theory, but rather constructed theories of our own during the process as the need arose. Our strategy has been largely an intuitive ability to devise the most effective way out of countless blind alleys and dead-ends, to arrive at certain rules and theories along the way.

These theories and computational methods have been evolving since 1988. We constantly improve on them as we fine-tune them and feed them more and more test sentences that turn up every day. Of course, they are still a far from perfection - even 80% perfection. Nonetheless, the methods used in ESI allow us to tweak, modify, enlarge, expand or restrict those rules to adjust the result to real language and thus keep slowly approaching the 80% goal, perhaps even a 90%, in the near future. This ongoing process

will probably never end, but with patience and continuous revision, will eventually take us to levels heretofore unattained.

Iii. What Makes ESI Different from other MT Systems?

If one compares ESI's performance with commercially available MT systems, such as Power Translator 7.0 or Systran Pro Version, the first impression might make one prematurely conclude that they are roughly equal, with ESI only slightly better.

This conclusion was actually the central issue of the Spanish magazine *PC Mania* in November 1999. *PC Mania* evaluated 5 programs, including Reverso Pro, Power Translator, ESI and others. According to that comparison, ESI came out first closely followed by Reverso Pro. The tests were done using English as the source language and translating various business, advertising, literature, science and news media texts.

Around the same time, *CALICO REVIEW* did another evaluation of ESI, this time comparing it with Power Translator and Systran. Below are some excerpts from the evaluation, which was physically carried by the linguistics department of Universidad La Coruña, in Spain. ESI ranked first, scoring 4.5 points out of a maximum of 5.

CALICO's Summary:

"English Spanish Interpreter is a useful tool for translation and composition due to its flexibility and reliability."...

"The program runs smoothly, is well-designed and provides the user with many tools, especially the English-Spanish dictionaries which are probably the best available at the moment in the price range of ESI."

"One of the most outstanding features of the program is the ability to display meanings or translations in tree format, thus providing a bird's eye view of the whole syntactic content and different meanings or translations of a term...

... and successfully handles any regular or irregular verb either in Spanish or English."

As indicated previously, it is in interactive mode that ESI works best."
"... when ESI is used interactively... the output can be improved up to almost human-like quality..."

"Altavista (Systran) ranks last, …It also fails to properly translate the verb "be" into Spanish "ser" or "estar", which the other two handle successfully.

"Power Translator would come second in accuracy as it fails, for instance, to properly handle noun adjective agreement in Spanish and the vocabulary is not always suitable..."

In the introductory paragraph in this section we mentioned the fact that ESI seemed to perform only slightly better than its competitors. At least, that was the reviewers' estimation, and we will not argue with them. However, many other factors will have to be taken into consideration in order to really understand the ESI engine's true power and do justice to its true worth.

With ESI Version 4.0 about to be released, much improvement has been made since CALICO's and PC Mania's evaluations. But it must be remembered that the evaluations were conducted merely from the standpoint of grammar. ESI has evolved in many other parallel fronts besides grammatical parsing, and it is precisely in those fronts that the most significant development has taken place. ESI should not be evaluated only on the basis of grammar and syntax, since it has many other important structural fortes. In order to grasp the concepts behind those fortes, we must reveal some of the foundational *secrets* within ESI's translation engine.

ESI operates through 12 stages, all of which are unique to Word Magic Software except of course, stages 1 and 6.

1. Sentence Parser
2. Style and Format Parser
3. Spell-checker and ambiguity-checker revision module
4. Idiomatic Construction Parser
5. Idiomatic Expression Parser (Dictionaries)
6. Grammatical Parser
7. Semantical Disambiguator
8. Field-Category Disambiguator
9. Pragmatic Disambiguator
10. Statistical Disambiguator
11. Translation Memory Module
12. Interactive Re-parsing module

Other modules include:

13. Grammar Display: Shows the different parts of the sentence in different colors
14. Interpreter Module: Translates the translation back into the User's own language

I shall briefly describe the main issues involved in each stage:

1) *Sentence Parsing*
The first step in any MT application should be the breaking down of the text into sentences and then into unit lexemes, that is, individual words. Not much is special about ESI's sentence parsing, except its ability to recognize true carriage returns and linefeeds, even when they are confused among artificially inserted ones that typically come in emails and HTML text formats. Moreover, it will not interpret as 'end of sentence' the periods that come in number figures, brands nor Internet addresses.

Another unique feature is ESI's ability to parse secondary sentences within the main sentence, when they are enclosed in parentheses or within hyphens. A sentence in parentheses is considered to be a comment and translated separately, without affecting the main structure in any way.

2) *Style and Format Parsing*
Although this problem has not been tackled by any other MT system that we know of, it is extremely important in obtaining a correct translation.

The grammatical rules that apply to a normal predicative sentence are not the same ones that apply, for instance, to a title, a heading, a list of terms, a glossary, a table, a display, an ad, an order or a message. So, it is important to tell the grammatical parser (stage 6) which kind of text is being fed in, before the parser starts to apply incorrect rules. Otherwise, results are guaranteed to be disastrous.

This is also the case with clauses (a, b, c, d, etc.) typical in legal documents. Each clause in itself may or may not be a complete sentence. It could also be part of the same predicate of the first clause. It is a difficult task to determine which rules to apply in each clause.

ESI is just starting to implement rules to deal with these complications, and we expect to have them in operation in version 5.0, by Fall 2003. In the meantime, the User can work around the problem by using the Interactive Mode under Options/Style.

Also, as a first approach to this powerful mechanism, ESI is implementing a full-blown word processor compatible with MS Word XP, which will automatically determine and reproduce format and style. This new word processor will come with version 4.0, slated to be formally released in March or early April 2003.

Besides sentence format, ESI also takes into account the format of each word. Capitalized letters are a constant source of problems, sometimes coming with first-letter uppercase and sometimes all uppercase.

In the first case, the word can be confused with a proper name. Take Ford, Smith, Carpenter or Book, for instance. There are thousands of those possibilities. Or take IBM, model AK-908, 128 M RAM, Dell, Logos, Super Translator, IN, Ca, CA, FLA, Fla. These are not included in dictionaries. Are they to be taken as first names, last names, product names, brands, company names, country names, zip codes, email addresses, abbreviations, or simply as words which the writer chose to capitalize in order to highlight what he wanted to say?

IN, for instance, could be a capitalized preposition or the abbreviation of Indiana (included in our dictionaries). AND could be the abbreviation of a substance, a Boolean (a noun) or a capitalized conjunction.

ESI has intelligent mechanisms for deciphering such ambiguities, even though much is still pending in this area.

Another type of formatting found in most texts is the quotation mark. It is almost impossible for MT to determine what exactly goes inside two sets of quotation marks. It could be an actual direct quote, a noun, a whole phrase or simply a device used to highlight a particular word or expression, or part of the main sentence. ESI has the capability of making a distinction and correctly translating

a. The word case,
b. The word “case”
c. The “word case”

3) *Spell-checker and ambiguity-checker revision module*

Spell-checking is necessarily a vital step in any translation process. ESI has two very powerful dictionaries for this purpose containing several million words and conjugated words in each language.

The Spanish spell checker includes all the possible combinations of enclitic verb forms. Version 5.0 will also include all diminutives, augmentatives and superlatives in plural, singular, and their feminine and masculine forms, whenever applicable.

The ambiguity checker is unique to ESI. It highlights all grammatically ambiguous words in the text for the User; that is, those which could be written either with or without an accent, meaning totally different things in each case.

In Version 5.0 this unique feature will be developed to a much more intelligent and user-friendly level, highlighting those words only where incorrect spelling is suspected.

Moreover, it will point out in English many common errors, such as using *it's* when *its* should have been used, or the incorrect placing of a

hyphen where it could be misinterpreted as *a phrase within hyphens* instead of a *hyphenated cluster of two or more words.* The translation of those two cases would produce radically different results. ESI is capable of treating both cases correctly, *provided the hyphen is properly placed.*

No other MT system that we know of is currently capable of correctly translating hyphenated clusters.

4) *Idiomatic-Construction Parsing*

We make a clear distinction between *idiomatic constructions* and idiomatic expressions.

An idiomatic expression is a word cluster that can be entered as an entry in a dictionary. The number of idiomatic expressions is considered limited, albeit very high and close to half a million in English alone. Idiomatic constructions, however, are in theory unlimited, since they are not set phrases but *non-orthodox syntactical patterns* used to construct phrases, admitting permutations of words. An idiomatic expression ceases to be one if we change one of its constituent words. For instance *"Let's call it a day"* ceases to be an idiomatic expression if we say instead *"Let's call it a microsecond"*. In other words, in the idiomatic expression, it is the *specific words* that form the set phrase. In idiomatic constructions, it is the particular pattern of clustering words that becomes idiomatic. Since the words themselves can change, the permutations can explode in number, and the only way to handle such exceptions is through an 'idiomatic parser'.

One way of creating idiomatic constructions is through the use of the hyphen. In the hyphenated cluster, one word becomes the operator of the other word, and the whole cluster takes on a part of speech and a meaning different from the orthodox free association. Take for instance: weak-sighted. We can change this cluster to strong-sighted, or far-sighted, near-sighted, feeble-minded, strong-chested, weak-chested, broad-shouldered, etc. We see that there is only one pattern from which many combinations can be formed.

Since no dictionary could possibly include all possible combinations, a system has to be devised to interpret these kinds of patterns. We do not know yet how many patterns there are, even in the case of hyphens. And hyphenation is not the only instance in idiomatic construction. There are hundreds of additional patterns. Here are a few examples of hyphenation:

strong-chested: adj + noun+ *ed => that has a strong chest, adj*
feeble-minded: adj + noun+ *ed => that has a feeble mind, adj*
feeble-mindedness: adj + noun+ *ed+ ness => weakness of mind, noun*
US-born : noun + participle => *born in the US, adj*

easy-to-use : adj + infinitive => *that is easy to use, adj*
chocolate-eating: noun + present participle => *that eats chocolate, adj*
science-fiction : noun + noun => noun

ESI has implemented a few cases of hyphenation, but not all, of course. This is another pending task of investigation. All other MT systems that we have tested failed in this task.

Another way of using hyphens is in the association of adjectives. Consider, for instance: "A red-plastic toy", and "A red plastic toy." ESI will correctly interpret and translate the first phrase as "A toy made of red plastic", whereas the second one will be "A plastic toy having red color."

Yet another use of hyphens is in verbs. Machine-dry, steam-clean, spray-paint, jump-start, etc. are all examples of hyphenated verbs. ESI deals with these cases through inclusion in the dictionary rather than treating them as special patterns. Here is sample sentence translated by ESI:

He dog-eared the pages using a cutting-edge technology.
ESI: *Él dobló la esquina de las páginas usando una tecnología de punta.*
Others: *Él perro-espigado las paginaciones usando una tecnología del corte-borde.*

Idiomatic constructions are also formed in English without the use of hyphens. Take, for instance, the normal, 'orthodox' sequence of adjectives in a noun phrase: aAA,A,...,AS, or aAAS, AAAS (Where, a = Article A=adjective S=noun D= Adverb V= Verb).

In other words, in English there is normally one or two series of the A before any S, separated or not by commas.

However, a particular idiomatic construction allows: aAASAS, or aASAS, or aAASS. Example:

A five year old boy. A twenty five foot long boat.

Translation by Word Magic: *Un niño de cinco años de edad. Un bote de veinticinco pies de largo.*

Translation by a typical MT system: *Un viejo muchacho de cinco años. Los veinte cinco pies desean barco.*

Notice that the above examples require the use of the hyphen. However, people leave hyphens and commas out in most cases nowadays, even though it is grammatically incorrect. A practical MT system has to be able to recognize such deviations from orthodox grammar, interpret them, and possibly point them out to the writer for correction.

Now let's look at another example, this time in an exclamational phrase:

What a nice car! (DaAS),

which seems "normal", but all existing programs on the market translate this as:

Qué un carro bonito! (DaSA)

whereas it should be translated as:

Qué carro tan bonito!: (DSDA)

There are hundreds of idiomatic constructions like these in Spanish, as well. ESI does not pretend to have identified them all, but it has actually identified a great number of them, and part of our goal is to discover, identify and solve more and more cases with every new ESI version.

Yet another type of Idiomatic Construction is one that comes from leaving out essential parts of the sentence. The example we presented before is a typical one: Leaving out the pronoun *I* from the sentence: *Hope she is we/I. Hope you come. Wish to see you soon.*

Here is another case where not one but several words are left out:

Mary will be five tomorrow.

Which should be interpreted as:

Mary will be five years old tomorrow.

As expected this, as well as hundreds of other idiomatic constructions, is not handled by any MT system. ESI, in most cases (but not all), interprets them correctly.

5) *Idiomatic Expressions*

There are thousands of idiomatic phrases in English, and the only way to account for them is by their inclusion into the database, since they do not follow any generalized pattern or rule. The same is true of Idiomatic Constructions.

An Idiomatic Expression is like a chemical compound: The final aggregate does not share the properties of any of its constituents but rather has its own unique properties.

After checking for Idiomatic Constructions and hyphenated clusters, ESI goes on to the next stage and identifies all possible Idiomatic Expressions contained in the sentence. This is done in close consultation with the Dictionary, which contains approximately 125,000 expressions as entries, with their corresponding equivalencies in the other language, for a total of 250,000 entries. This is by far the largest collection of idiomatic expressions and clusters in computer-dictionary form existing to date.

However, we must note that idiomatic expressions or set-phrases cannot and should not be taken at face value whenever they appear in the text and the dictionary. A Parsing System is necessary. Word clusters have a grammar of their own, and special *cluster-grammar rules* have to be implemented in order to make a sensible interpretation of either English or Spanish texts. Consider the set-phrase:

Enough of it! = ¡Basta ya! (Enough is enough!)

If we say: *"We had enough of it"*, the idiomatic sense outlined above is totally out of context and the dictionary entry should be rejected, making instead a literal translation. Otherwise it would be translated into Spanish as: *We had enough is enough!* ESI is capable of making such distinctions and the translation "*basta ya"* is not picked up by the Idiomatic Parser. Or, consider these two other examples involving nouns:

a) The system controls = Los mandos del sistema (the control panel)
b) The system controls our economy= El sistema controla nuestra economía

In this case, *"system controls"* exists as an entry in ESI's dictionary, but it had to be rejected from the selection because it is out of context.

Perhaps these are rare instances of word clusters being rejected. However, with adverbs, verbs and adjectives the probability of an out-of-context idiomatic expression is fairly common. These exceptions will multiply as ESI's database adds more and more idiomatic entries.

With verbs, the selection of idiomatic expressions taken from existing entries in the dictionary is a very large and complicated process. In fact, it is the single largest section in the whole program, and it includes hundreds of different rules which take into account transitivity, reflexiveness, semantic attributes and the verb's virtual environment. For instance, if you enter: *"I will **ask** Jane **out**"*, ESI will translate it as *"Invitaré a salir a Jane." (I will invite Jane to go out)*

However, if you enter *"I will ask the car out"*, ESI will not accept the dictionary entry "ask out". It will render a literal translation which is meaningless both in Spanish and in English.

We have tested these sentences with other MT systems, and they were not able to recognize the idiomatic expression in either case, rendering:
"Pediré Jane hacia fuera. (I will ask Jane towards the outside)."
Here is a trickier verb, *"**turn on**":*

*She **turned** the switch **on** = Ella encendió el interruptor.*
*She **turned** him **on** = Ella le excitó.*

*She **turned on** him = Ella se volvió contra él*
*She **turned on** the corner = Ella cambió de dirección {dobló} en la esquina.*

See below a comparative translation made by the popular Systran program:

Ella giró el interruptor
Ella lo giró
Ella lo giró
Ella giró la esquina.

Or, consider the following idioms:

*I will **take** my son **to** his new room = Llevaré a mi hijo a su cuarto nuevo.*
*I will **get** my son **to** do his homework = Persaudiré a mi hijo a hacer su tarea.*
*She **came to** my house = Ella vino a mi casa.*
*She **came to** last night= Ella recobró el conocimiento anoche.*
*She **came to** work=Ella llegó a trabajar.*
*The smith **worked** the metal piece **in** very skillfully= El herrero insertó el pedazo de metal muy diestramente.*
*The smith **worked in** our car = El herrero trabajó en nuestro coche.*

As we see through these examples, the selection of phrasal verbs is not as simple as reading them in the text and comparing them to their match in the database. There are numerous rules involved, all of which had to be developed by Word Magic. We cannot discuss all the possibilities that arise when trying to fit a phrasal verb into a given text, but this last example should suffice to point up the complexity involved.

Let's consider the verb **have to**, which is a paraphrastic case that gives the phrasal compound the connotation of an obligation to do something, as in "I **have to** go to school."

In the sentence: "I **have to** go to the party", the cluster is first recognized as existing as an entry in ESI's dictionary and then, after going through the Idiomatic Parser, it is finally accepted with its corresponding translation "**tener que**", and the final translation would be.

Tengo que ir a la fiesta.

However, if ESI encounters exactly the same cluster **have to** in this other sentence

This is the only dress I **have to** go to the party.

It will not interpret "have to" as a word cluster, even though it exists as an entry in the dictionary, but instead it will split the sentence right in the middle of **have to:**

[This is the only dress **that I have]** + [**to** go to the party]
And its translation
Este es el único vestido **que** yo tengo] para ir a la fiesta

Notice that a conjunction "**que**" has been added to the Spanish rendition, which was omitted in English.

6) Grammatical Parsing

In this area, ESI is similar to other MT systems in that it **must** have a way of parsing the syntax of the input text based on fixed, orthodox grammatical rules. Therefore, we will not delve into the typical *grammatical constructions,* but rather highlight those which we know are not handled at all by any other translator available on the market.

In particular, ESI is the only system capable of detecting the elision of the relative pronoun **that,** which usually precedes a subordinate or relative phrase, and it accomplishes this on the basis of grammatical criteria alone.

The following example illustrates what we mean:

The Internet is a technology people use around the world.

There is *silent* relative pronoun somewhere, hidden in the syntactic construction of that sentence.

If we try Systran or any other program available on the Internet, the parsing will be: SVaSSSraS, that is:

Internet is a usage of technology-people around the world,
Translated into Spanish as
El Internet es un uso de la gente de la tecnología alrededor del mundo.
However, ESI correctly interprets this construction as: SvaS + that + SvraS, or

*The Internet is a technology **that** people are using around the world,*
translated as:

Internet es una tecnología que las personas usan alrededor del mundo.

Other constructions which are generally incorrectly handled by most other applications are long strings of nouns, proper names, numbers, and certain special constructions. We invite the reader to test and compare these cases through our online applications at our web site http:// www.wordmagicsoft.com/ and compare.

There is one final point that we must stress with respect to ESI's interpretive power. The above example should serve to illustrate that adding just one more degree of freedom to a syntactical interpretation, in order to be able to handle elided relative pronouns linking relative clauses, could double, or perhaps more than double, the possibilities of a faulty interpretation. Permutations increase exponentially with the number of degrees of freedom in any closed system. And ESI, as we have seen, has not one more but *many more* degrees of freedom in its code, to be able to analyze, interpret and accept not only diverse grammatical structures, but also idiomatic expressions, hyphenation, parentheses and idiomatic constructions.

Additionally, consider one other factor: ESI's dictionaries are much larger than any other digitalized translation dictionary available. Therefore, the possible number of permutations is tremendously larger too, as we saw at the beginning of this article. Dealing with 70,000 words is not nearly as complex as dealing with 250,000 and 850,000 translations references in all, as ESI does. The possibilities are endless - but then, the possibilities of problems and errors are endless too.

Perhaps a comparison with real-life systems will shed light on this situation. ESI is like a jetliner flying at high speed with hundreds of people on board, whereas other primitive MT systems are more like hand-driven, four-wheeled wagons riding on rails.

The wagon has practically no degrees of freedom: It can only move straight ahead. It is restricted vertically by the ground, and horizontally by fixed rails.

The airplane can move up and down, laterally, nose-up, nose-down, sway, tilt over to right or the left, glide, or even plummet down out of control. It can encounter strong winds in its flight, snow, sleet, rain and storms. It also has to be able to glissade through gentle breezes and then land softly on the ground.

No doubt, to construct such a machine and to succeed in its stabilization and maneuverability is much tougher than doing the same with the railroad wagon. Also, it would take much more time to achieve the desired stability, after traversing a longer road of experimentation through trial and error.

7) *Semantical Disambiguation*

The theory and practice we have implemented in ESI - particularly the practice - deviates from what Noam Chomsky teaches that the syntax of a sentence should be determined independently from its semantics.

> *Syntax was regarded as the heart of linguistics and (Chomsky's) project was supposed to transform linguistics into a rigorous science.*
> John R. Searle End of the Revolution

We have found through thousands of instances taken from real-life texts that syntax is totally dependent on semantics, and ESI's grammatical parser is equipped with a corresponding number of rules and exceptions to deal with this reality. This dependency is even more dramatic in Spanish. Unfortunately, space does not permit in this article to explore in depth this fascinating new field, a field which we believe is being researched and developed for the first time ever by Word Magic Software with plenty of success.
We will ponder just a few examples. Look at the following two sentences

a. My daughter is to marry tomorrow = Mi hija ha de casarse mañana.
b. My idea is to marry tomorrow = Mi idea es casar**me** mañana

ESI's rendition in each case is radically different from the other, as you can see at the right. The typical translations found in other MT applications, on the other hand, make little or no sense, because none of them, with the sole exception of ESI, as far as we know, take semantics into consideration.

The trick here is to recognize that an idea cannot get married (case b), and therefore the action is passed on to whoever is writing the sentence, whereas in case a. the daughter is the one that explicitly is to get married.

Here is another example where semantics plays a key role. Consider the next two sentences:

a. Wine from Europe and cheese in general will be affected by the tax.
b. Wine from Europe and Asia in general will be affected by the tax.

Here are ESI's translations:

a. El vino de Europa y el queso en general **serán afectados** por el impuesto.
b. El vino de Europa y Asia en general **será afectado** por el impuesto.

Notice that ESI interprets the subject of sentence a. as a noun phrase composed of two elements (wine and cheese), and thus uses the plural form of the verb and its predicate, whereas in sentence b. ESI correctly recognizes one element only and thus uses the singular verb.

Disambiguation properly occurs when ESI chooses among several connotations of a noun, an adjective or a verb based entirely on their mutual

semantic correspondence. In Chapter I we presented an example of this process:

I am teaching this bat to fly,

where ESI correctly selected the meaning 'murciélago' (animal) for the word **bat**.
The process can operate the other way around too: Consider the following sentence: *This bat is used to play baseball = Este bate se usa para jugar béisbol.*

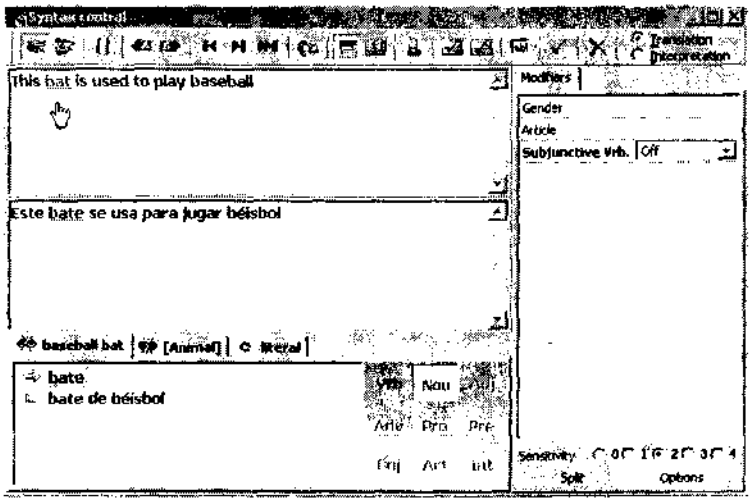

Fig.2. Interactive Window showing default translation of sentence after the noun 'bat' is selected. Notice vertical display of translations and horizontal display of meaning-tabs.

Note that ESI is choosing **bat** as *baseball bat* by default, as a first choice. What would happen if we *force* ESI to choose bat as an animal? We can do this accessing the Interactive Module, shown at the left, highlighting *bat* and then clicking on the tab 'animal'. Results are shown in Fig. 3, below.

We see now that ESI renders: *Este murciélago* ***está acostumbrado a*** *jugar béisbol*, where the verb has actually been changed to better fit the correspondence with animal, that is: ***be accustomed to***

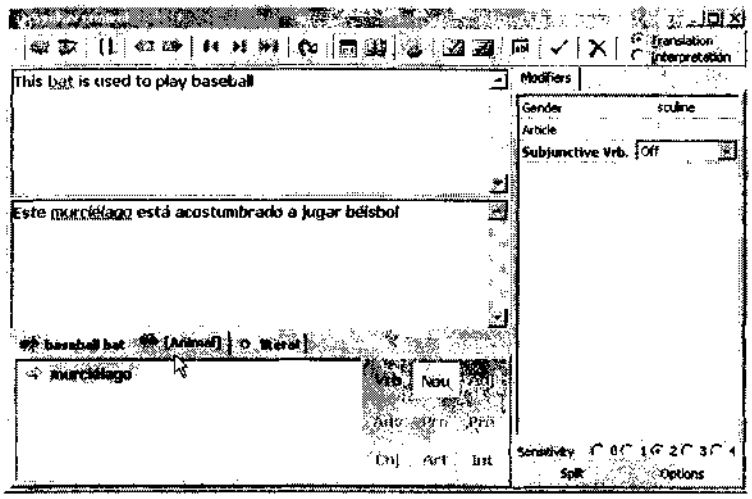

Fig.3. Interactive Window showing user-selected translation after clicking on the tab 'animal'. Notice live interaction of other parts of the sentence, in this case, the phrasal verb 'be used to'.

The equivalence in English would now be: *This bat (animal) is accustomed to playing baseball.*

There are hundreds of other instances where semantics is used to perform noun disambiguation, adjective disambiguation, verb disambiguation, or all of them simultaneously.

8) *Field Category Disambiguation*

ESI relies on hundreds of thousands of semantic marks, field-category attributes and pragmatic-field attributes embedded in its database. Category attributes are classified in two groups: industry-specific and media-specific categories. The first group includes almost 200 categories, such as Medicine, Computers, Internet, Law, Business, Accounting, Forestry, Agriculture, and Linguistics.

The second group includes specific tools or media used in each main category and includes some 400 different items. One category-one item and one category-two item are assigned to each entry and each meaning in the dictionary. Nouns, adjectives and verbs alike are marked, so as to provide a wide-base correspondence. In addition, you can specify any particular category as your preferred field for a particular document, and the automatic output will correspond to the most appropriate or the closest

meaning to the specified field. This functionality is currently not available but will be incorporated in Version 5.0

9) *Pragmatic Disambiguation*
Similarly, a collection of 500 pragmatic attributes are being entered into ESI's database. Each entry in the dictionary can hold up to 18 different attributes in each part of speech and each meaning. This finer disambiguation system is still not operative in English to Spanish parsing, but we found it to be necessary during the development of Spanish to English.

Space does not permit us to delve further into this field, but a few examples should suffice to point up the necessity of implementing this extensive markup system, and also the immense complexity of the tasks involved in Spanish interpretation.

Consider the phrase: *Se quebró el vaso.* This is translated into English as: *The glass broke.*

However, a syntactically identical phrase such as: *Se quebró la nariz, is* translated in totally different form: ***He*** *broke* ***his*** *nose.* A tacit-subject (omitted from the text) condition applies in the latter case. Why? Only because *we know* it is so: A pragmatic inference.

Another example: *Me empujó la espalda.* Again, this is a tacit-subject situation where we assume that there is somebody out there - not mentioned in the text - who is performing the action. ***He*** *pushed my back.* The same would apply to:

Me empujó el carro = **He** pushed **my** car.
Me compró un regalo = **He** bought **me** a gift* .
(*Notice that in this case it is **not**: He bought **my** gift)

However, if we now say

Me empujó el viento = **The wind** pushed me,

we notice that the tacit pronoun mysteriously disappears! We now have a *reverse-subject* case, that is, a case where the subject *(the wind)* is placed behind the verb, rather than in front of it.

The only way to know when we have a tacit pronoun or a reverse-subject is by simple, straightforward *guesswork*, or by what academically is known as *pragmatic inference.*

A final example will show us that this happens not only at the beginning of sentences. It can happen anywhere within the syntactic structure. Here is a case where we are confronted with a *pragmatic inference* in a relative clause:

*a. El carro que tuvo el accidente = The car that **had** the accident.*
*b. El accidente que tuvo el carro = The accident that the car **had.***

As we can see, the two Spanish sentences are constructed identically from a syntactic point of view. However, their respective translations are different. Case a. places the verb *(had)* in front of the relative complement *(accident),* whereas case b. requires placing the verb at the end of the sentence, because it so happens that car is the subject in both cases.

ESI correctly treats correctly these and many other similar cases in Spanish. In English there are many more applications of this *pragmatic markup* scheme, but they will be implemented in version 5.0

10) *Statistical Disambiguator*

Even with all the millions of bits and bytes of information included in Word Magic's databases, with semantic attributes available as part of each dictionary entry, category and pragmatic fields and more than 250,000 word clusters and idiomatic expressions, many times there is not enough information available to make a proper disambiguation. Sometimes it cannot be determined if a word that ends in ***ed*** is a adjective, a past-participle or verb in past tense. Sometimes it cannot be determined if a word ending in ***ing*** is an adjective, a present participle, a verb in progressive mode or a true gerund.

In those cases ESI has to resort to statistical usage information. However, the statistical approach used in ESI is not the same as the one we discussed at the beginning of this article. Each word or word cluster in ESI's dictionary has a usage level which indicates the probability of its use in daily speech.

Common, daily-used words are marked as level one. Normal words used typically at high-school and university levels are marked as level two. Difficult, rarely-used words are marked as level three. Level four is for normal technical, industry-specific terms. Level five is reserved for technical words which could be confused with other ordinary words, thus necessitating special labeling.

Thus, in those cases where nothing else provides the signaling for disambiguating a sentence, the lowest usage-level combination is selected, either for a single or a word cluster.

11) *Translation Memory*

With this new feature, available in the soon-to-be-released Version 4.0, Word Magic Software is seeking to implement the best of two worlds: the world of Machine Translation (MT) with a most practical application of

Translation Memory (**TM**), that is, the ability to pre-select the first translation of a particular word from one language to another.

We represent this with the formula: **MT + TM = ESI**, a fresh concept introduced to the market for the first time ever in our constant endeavour to produce the ideal machine for English-Spanish translation and interpretation.

ESI's TM system is designed to operate in both languages. When translating from English to Spanish you can select and mark any word in Spanish as the first translation. When translating from Spanish to English you can select and mark any word in English as the first translation, which might not necessarily be the reverse of the Spanish choice.

12) *Interactive Mode*

Even though ESI is a sum total, a composite of numerous innovative and diverse methods in MT, perhaps the most unique and most conspicuous is the Interactive Mode. In this mode the User interacts with the translation of the document sentence by sentence - rather than allowing a complete automatic translation - and has the ability to change any part of speech, any meaning and any synonym of any word forming the sentence. The most striking feature of this interaction with the MT system is that ESI is capable of processing all these changes **live,** thus giving the User instant feedback on the results of the choices he or she makes. Moreover, each change affects all other words of the sentence, producing all the necessary adjustments in number, gender and semantics to render another coherent translation, perhaps with a totally different meaning.

We saw an example of this interrelation with the case *"This bat is used to play baseball.* When we changed the selection of *bat* from baseball bat to animal, the system also automatically changed the verb *"be used to"* from *"be utilized to"* to *"be accustomed to",* thus rendering a final output more in line with what it is expected from an animal.

Another example - among hundreds of others* that we could discuss if we had the space in this article - is in the formation of the Spanish subjunctive. Let's consider the sentence:

a. I need the boy to go to the store = Necesito que el chico vaya a la tienda.
b. I need the boy to go to the store = Necesito al chico para ir a la tienda.

Sentence a. is subjunctive by default. You can change it to b. just by clicking on a button on the Modifiers Panel in Interactive mode. (See pointing hand in Fig. 5)

We can also do the opposite. We can start with a sentence which is not subjunctive by default, such as *"I need the car to go to the store"*, and see what happens when we click on the same button.

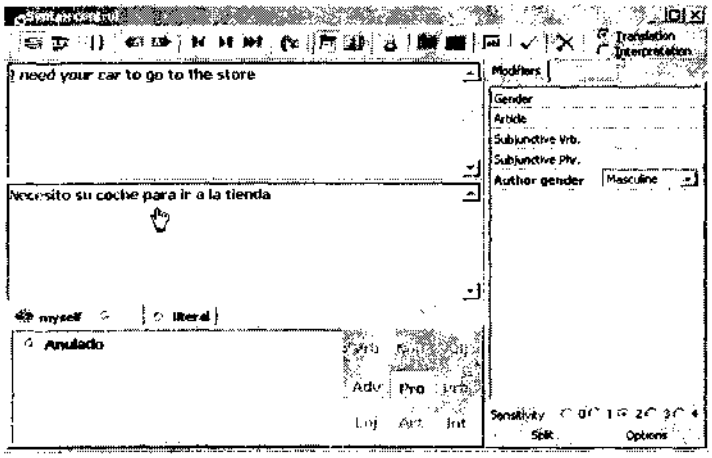

Fig 4. Changing sentence to subjunctive mode the Modifiers Panel.

ESI will automatically determine that the sentence shown at the left is not to be translated into Spanish in the subjunctive mode.

However it offers you the possibility by activating the Subjunctive-Phrase selector button. See pointing hand in Fig. 5

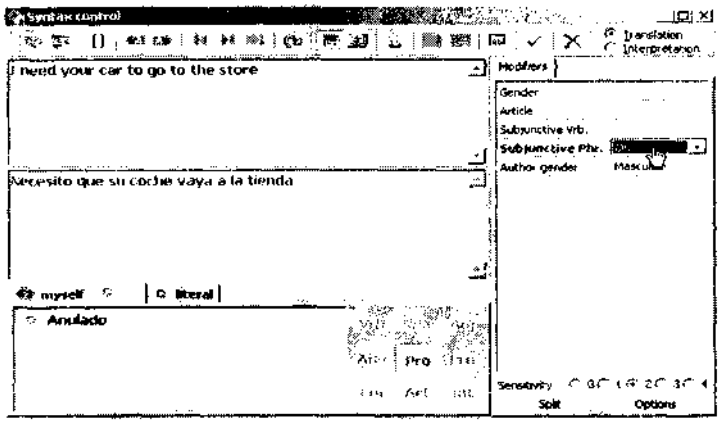

Fig 5. Forcing a sentence to be translated in Spanish subjunctive by clicking on the Subjunctive-Phrase button, above. (Pointing hand)

The Modifiers Panel, shown at the far-right of this image displays different options triggered by the ambiguities presented by the sentence at hand. In this instance, ESI detected the possibility of a subjunctive case and thus, it displayed the respective button. It activated also four other grammatical options.1

CONCLUSION

We have observed that ESI combines different methods in its engine. If we had to summarize which are the most significant features that set ESI apart from other typical MT English-Spanish translation applications available today, perhaps we should point out the following:

1. Word Magic Software has its own customized dictionaries which besides being massive, (larger than its nearest competitor perhaps by a factor often), include a comprehensive markup system in the database providing millions of bytes of detailed information about the characteristics of each word. The dictionaries, besides being customizable are constantly growing with each successive version.
2. ESI's engine relies not only on grammar and syntax; it is the only engine that bases its logic on context, semantics, field-categories and pragmatic attributes taken from a '*knowledge of the world database'*.
3. ESI is the only MT system that offers a truly live Interactive Mode, one capable of displaying instant feedback and making all necessary grammatical and other adjustments simultaneously.
4. ESI's dictionaries have a collection of 250,000 idiomatic expressions, including their translations. Thus, ESI is the first MT system which attempts to translate the meaning from one language to another while conserving the idiomatic nuances. Its goal is not only to break the communication barrier, but the cultural barrier as well.
5. ESI is truly a productivity tool for the professional translator.

Isaac Asimov rightly stated that *"There is no distinguishable difference between magic and high-technology in the eye of the beholder."* Our goal is to keep improving ESI and TDT until we can indeed claim that we are displaying '*magic in the eye of the beholder'*.

NOTES

1 You can see many more examples of the Interactive Mode by downloading ESI from our website, and going through the free Tutorial.

RICARDO ARGÜELLO MÁNTICA
Word Magic Software Corporation
www.wordmagicsoft.com
Email: Ricardo@wordmagicsoft.com