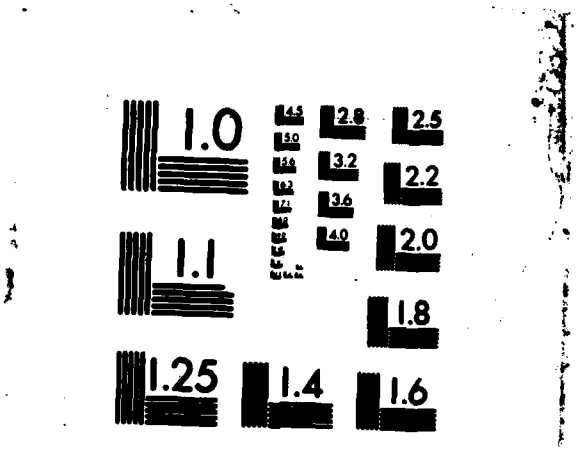
AD-A193 631 UNITRAN: AN INTERLINGUAL MACHINE TRANSLATION SYSTEM(U) 1/1
MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL
INTELLIGENCE LAB B J DORR DEC 87 AI-M-998
UNCLASSIFIED N00014-80-C-0505 F/G 5/7 NL



(4)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AIM-998	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): UNITRAN: An Interlingual Machine Translation System	5. TYPE OF REPORT & PERIOD COVERED: AI Memo 9/84-5/87
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Bonnie Jean Dorr	8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0505 (ARPA-ONR) N00014-85-K-0124 (ARPA-ONR) DCR-85552543 (NSF-PYI)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: December, 1987 13. NUMBER OF PAGES: 13 (including cover)
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Information Systems, Arlington, VA 22217	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

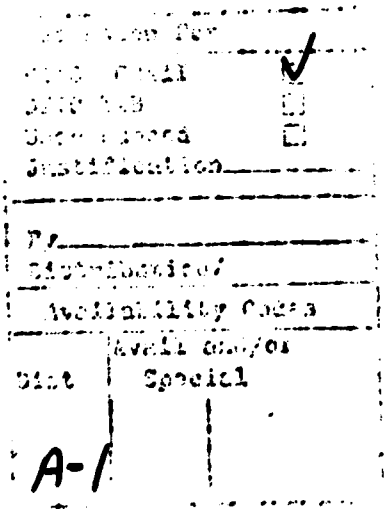
natural language processing, interlingual machine translation, principles and parameters, parsing, generation, linguistic constraints

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See back.

Number 20.

Machine translation has been a particularly difficult problem in the area of Natural Language Processing for over two decades. Current translation systems either have limited linguistic coverage, or they have poor performance due to formidable grammar size. This report presents an implementation of an alternative approach to natural language translation. The UNITRAN (UNIversal TRANslator) system relies on principle-based descriptions of grammar rather than rule-oriented descriptions. The approach taken is "interlingual", *i.e.*, the model is based on universal *principles* that hold across all languages; the distinctions among languages are then handled by settings of *parameters* associated with the universal principles. Interaction effects of linguistic principles are handled by the system so that the programmer does not need to specifically spell out the details of rule applications. Only a small set of principles covers all languages; thus, the unmanageable grammar size of alternative approaches is no longer a problem.



(A

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 998 **December, 1987**

UNITRAN: AN INTERLINGUAL MACHINE TRANSLATION SYSTEM

Bonnie J. Dorr

((Universal Translator)

ABSTRACT:

Machine translation has been a particularly difficult problem in the area of Natural Language Processing for over two decades. Early approaches to translation failed, partly because interaction effects of complex phenomena made translation appear to be unmanageable. Later approaches to the problem have been more successful but are based on many language-specific rules of a context-free nature. To try to capture all of the phenomena allowed in natural languages, context-free rule-based systems require an overwhelming number of rules; thus, such translation systems either have limited linguistic coverage, or they have poor performance due to formidable grammar size. This report presents an implementation of an alternative approach to natural language translation. The UNITRAN system relies on principle-based descriptions of grammar rather than rule-oriented descriptions. The approach taken is "interlingual", *i.e.*, the model is based on universal *principles* that hold across all languages; the distinctions among languages are then handled by settings of *parameters* associated with the universal principles. The grammar is viewed as a modular system of principles rather than a large set of *ad hoc* language-specific rules. Interaction effects of linguistic principles are handled by the system so that the programmer does not need to specifically spell out the details of rule applications. Only a small set of principles covers all languages; thus, the unmanageable grammar size of alternative approaches is no longer a problem. ←

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Laboratory's artificial intelligence research has been provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contracts N00014-80-C-0505 and N00014-85-K-0124, and also in part by NSF Grant DCR-85552543 under a Presidential Young Investigator's Award to Professor Robert C. Berwick. Useful guidance and commentary were provided by Ed Barton, Bob Berwick, Bruce Dawson, and Sandiway Fong. This report is an extended version of a paper that is in the Proceedings of the Sixth National Conference on Artificial Intelligence (1987).

[1]The name UNITRAN stands for UNIversal TRANslator, that is, the system serves as the basis for translation across a variety of languages, not just two languages or a family of languages.

Phenomenon	*Source Language Sentence*	*Target Language Sentence*
Verb Preposing	¿Qué vio Juan?	What did John see?
Null Subject	Vio al hombre.	{He, She} saw the man.
Clitic Doubling	Juan lo vio al hombre.	John saw the man.
Subject-Aux Inversion	Has John seen the man?	¿Ha visto Juan al hombre?
Embedded Clauses	The man that John saw that ate dinner left.	El hombre a quién Juan vio que co la cena salió.

Figure 1: The phenomena handled by UNITRAN include Verb Preposing, Null Subject, Clitic Doubling, Subject-Aux Inversion and Embedded Clauses. These phenomena are instrumental in understanding the parametric variations between Spanish and English.

1 Introduction

The problem addressed in this report is the construction of a translation model that operates cross-linguistically without relying on language-specific context-free rules. Many machine translation systems are non-interlingual approaches that depend heavily on context-free rule-based systems. For example, Slocum's METAL system (1984, 1985) developed at the Linguistics Research Center at the University of Texas is a transfer approach that relies on numerous language-specific context-free rules per language, solely for syntactic processing. The aim of this report is to present the computational framework for UNITRAN, a syntactic translation system currently operating bidirectionally between Spanish and English, and to put into perspective how the design of the system differs and compares to other translation designs. The distinction between rule-based (non-interlingual) and principle-based (interlingual) systems will be presented, and the advantages of the principle-based design over rule-based designs will be discussed. Finally, an overview of the UNITRAN design will be given, and a translation example will be shown.

The model that has been constructed is based on abstract principles initially set forth by Chomsky (1981) and several other researchers working within the "Government and Binding" (GB) framework. The grammar is viewed as a modular system of principles rather than a large set of *ad hoc* language-specific rules. Several types of phenomena are handled without sacrificing cross-linguistic application and without relying on a large set of language-specific rules. (Some examples of the phenomena handled by the system are in figure 1.) The system is designed so that the grammar-writer has access to parameter settings, thus enabling additional languages to be handled by the system. Before the source language processing (parsing) takes place, the parameters are set according to the source language values specified by the grammar-writer, and are then *reset* according to the target language values specified by the grammar-writer before target language processing (generation) occurs. For example, there is a "constituent order" parameter associated with a universal principle that requires there to be a language-dependent ordering of constituents with respect to a phrase. This parameter is modifiable by the grammar-writer, who sets the parameter to be *head-initial* for a language like English, but *head-final* for a language like Japanese.

Translation in UNITRAN is primarily syntactic; thus, there is no global contextual "un-

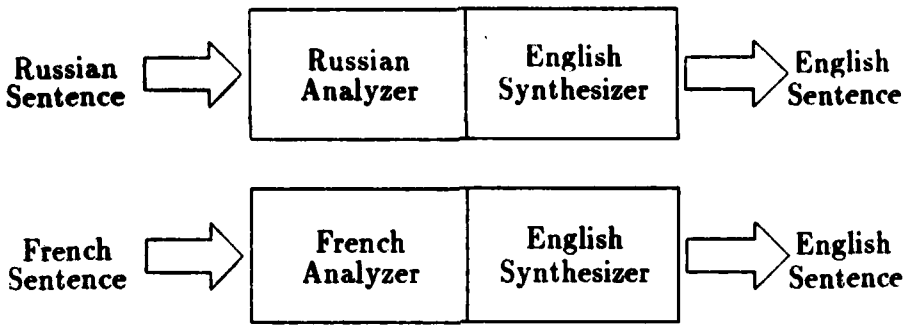


Figure 2: The direct approach, as found in GAT (1964), is a word-for-word translation scheme in which there is an analyzer and synthesizer for each source-target language pair.

derstanding" (the system translates each sentence in isolation). Semantics is incorporated only to the extent of locating possible antecedents of pronouns (*e.g.*, linking *himself* with *he* in the sentence *he dressed himself*), and assigning semantic roles (*e.g.*, designating *he* as "agent-of-action" in *he ate dinner*) to certain elements of the sentence, in particular, arguments of verbs (*e.g.*, in the English sentence "I read the book", the external argument (agent) of *read* is *I*, and the internal argument (theme) is *book*).[2] It should be noted that determining the mapping between semantically equivalent verbs is not a trivial task. For example, although the Spanish verb *gustar* is semantically equivalent to the English verb *like*, the argument structures of these two verbs are not identical. The subject of the verb *like* is the *agent*, whereas the object of the verb *gustar* is the *agent*. In order to include such cases of thematic divergence, the argument structure of a source language verb must be matched with the argument structure of the corresponding target language verb before substitution takes place.

The next section describes early (rule-based) approaches to translation. These non-interlingual systems will be compared to the interlingual (principle-based) design of UNITRAN and other systems in subsequent sections.

2 Direct and Transfer Approaches: Rule-based Systems

An early approach to translation (*e.g.*, the Georgetown Automatic Translation system (1964)) was a *direct* word-for-word scheme in which there was an analyzer and synthesizer for each source-target language pair (see figure 2). The primary characteristic of such an approach is that it is designed to translate out of one specific language into another.

Later approaches to translation (*e.g.*, the METAL system by Slocum (1984)) have taken a *transfer* approach, in which there is only one analyzer and one synthesizer for each source and target language. In this approach, there is a set of *transfer* components, one for each source-target language pair (see figure 3). The transfer phase is actually a third translation

[2]This is not to say that semantic issues should be ignored in machine translation; on the contrary, semantics may be the next step in the evolution of the translation system presented here.

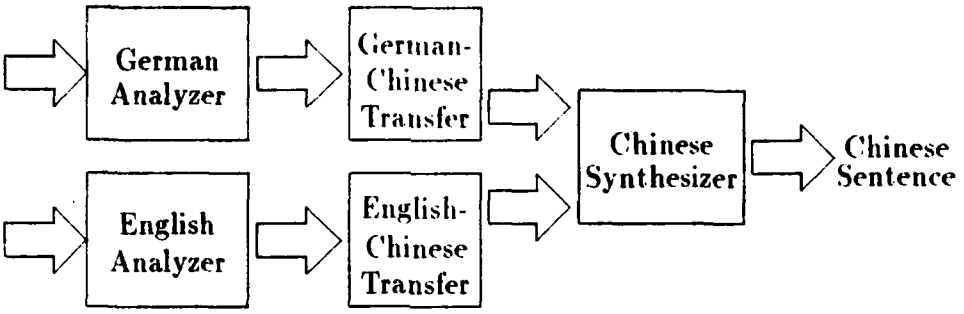


Figure 3: The transfer approach as found in METAL (1984) makes use of a set of transfer components, one for each source-target language pair. The source language sentence is first mapped into a source transfer form. This form is then mapped to a target transfer form that is used to generate the target language sentence.

stage in which one language-specific representation is mapped into another. In contrast to the direct approach to translation, the transfer approach has been somewhat more successful, accommodating a variety of linguistic strategies across different languages. The METAL system currently translates from German into Chinese and Spanish, as well as from English into German.

The malady of the transfer approach is that each analysis component is based on language-specific context-free rules. In Slocum's system, the type of grammar formalism is allowed to vary from language to language; however, regardless of the type of grammar formalism employed, each analyzer is nevertheless based on a large database of rules of a context-free nature. For example, the German analyzer is based on phrase-structure grammar, augmented by procedures for transformations, and the English analyzer employs a modified GPSG approach. (See Gazdar *et. al.*, 1985). Because the system has no access to universal principles, there is no consistency across the components; thus, each analyzer has an independent theoretical and engineering basis. Rather than abstracting principles that are common to all languages into separate modules that are activated during translation of any language, each analyzer must independently include all of the information required to translate that language, whether or not the information is universal. For example, agreement information must be encoded into each rule in the METAL system; there is no separate agreement module that can apply to other rules. Consequently, in order to account for a wide range of phenomena, thousands of idiosyncratic rules are required for each language, thus increasing grammar search time. Furthermore, there is no "rule-sharing" — all rules are language-dependent and cannot apply across several languages.

3 Interlingual Approaches: Principle-based Systems

The translation model described in this report moves away from the language-specific rule-based design, and moves toward a linguistically motivated principle-based design. The approach is *interlingual*, (*i.e.*, the source language is mapped into a form that is independent of any language); thus, there are no transfer modules or language-specific context-free rules.

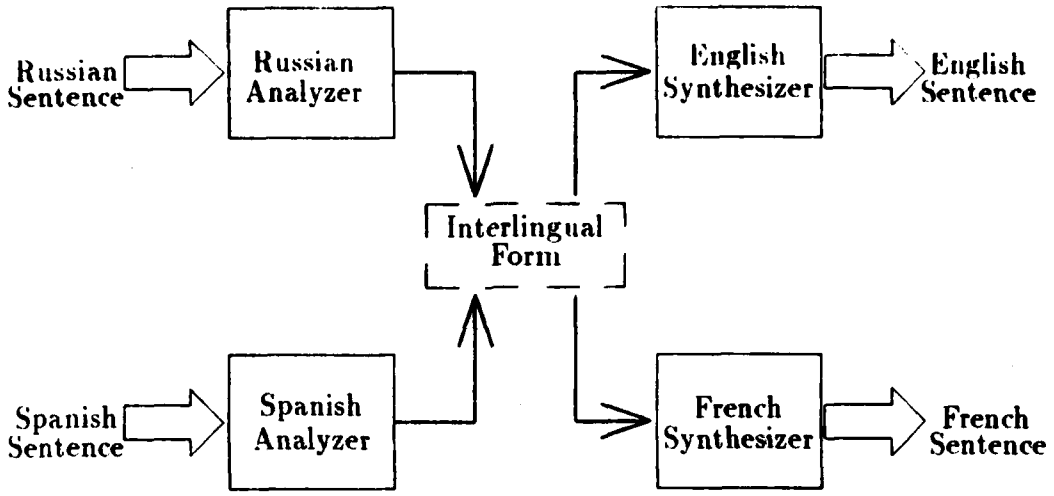


Figure 4: The interlingual approach taken by CETA (1961) and Sharp (1985) eliminates the need for a transfer component by providing a common underlying form. However, a separate analyzer and synthesizer is needed for each language. Also, because language-specific mechanisms are used by both systems, the grammar-writer cannot easily add new languages to the system.

Before describing the interlingual design as embodied by the UNITRAN system, we will first examine the interlingual design of the CETA and Sharp systems.

3.1 Interlingual Design: CETA and Sharp

The interlingual approach to translation has been taken by CETA (Centre d'Etudes pour la Traduction Automatique),[3] and Sharp (1985). However, the CETA system is not entirely interlingual since there is a transfer component (at the lexical level) that maps from one language-specific lexical representation to another. Sharp's system, although not rule-based, is also not entirely interlingual since context-free rules (set up for English-like languages) are hardwired into the code rather than generated on the fly using linguistically motivated principles; thus, languages (like German or Japanese) that do not have the same order of constituents as English cannot be handled by the system. The result is that the class of languages that can be translated is limited. The interlingual approach as embodied by CETA and Sharp is illustrated in figure 4. Note that there are no transfer components, but that there is a separate analyzer and synthesizer for each source and target language. The interlingual form is assumed to be a form common to all languages.

There are two problems with this incarnation of the interlingual approach. First, the grammar-writer must supply an analyzer for each source language and a synthesizer for each target language. Second, the grammar-writer has limited access to the parameters of the system. For example, the "constituent order" parameter mentioned in section 1 is not available for modification in the interlingual approach as embodied by CETA and Sharp.

[3]Grenoble University, France, 1961.

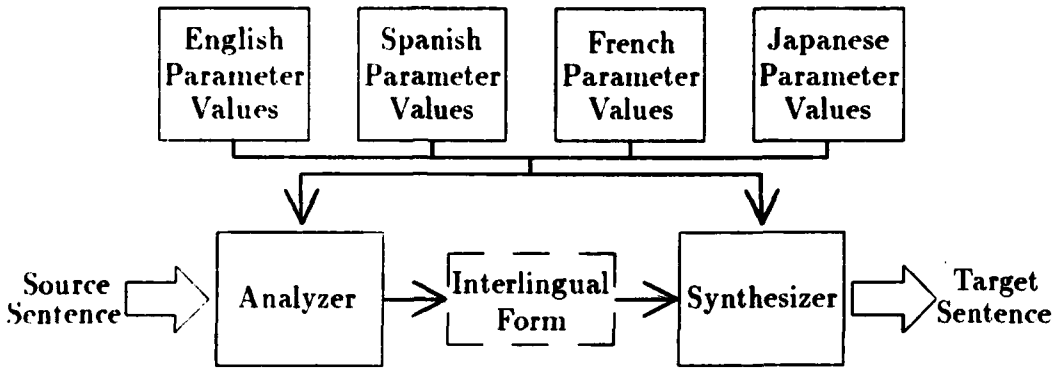


Figure 5: The interlingual approach taken by Dorr in UNITRAN uses the same analyzer and synthesizer for all languages. The grammar-writer may modify the parameters of the system in order to handle additional languages.

3.2 Interlingual Design: Dorr

The approach taken for UNITRAN is still interlingual by definition (*i.e.*, the source language is mapped into a form that is independent of any language), but the design is slightly different from that of CETA and Sharp: the same analyzer and synthesizer are used for all languages. Furthermore, the grammar-writer is allowed to specify parameter values to the principles, thus modifying the effect of the principles from language to language. This more closely approximates a true universal approach since the principles that apply across all languages are entirely separate from the language-specific characteristics expressed by parameter settings.[4] Figure 5 illustrates the design of the model. The analyzer and synthesizer are programmable: all of the principles associated with the system are associated with parameters that are set by the grammar-writer. Thus, the grammar-writer does not need to supply a source language analyzer or a target language synthesizer since these are already part of the translation system. The only requirement is that the built-in analyzer and synthesizer be *programmed* (via parameter settings) to process the source and target languages. For example, the grammar-writer must specify that an English sentence requires a subject, but that a Spanish sentence does *not* require a subject. This is done by setting the "null subject" parameter to TRUE in Spanish; by contrast, this parameter must be set to FALSE for English. (For details on the null subject parameter, see van Riemsdijk and Williams, pp. 298-303.) A dictionary for each language must also be supplied.

The translation system consists of three stages: First, the parser takes a morphologically analyzed input and returns a tree structure that encodes structural relations among elements of source language sentence. (This structure is the "interlingual" representation that underlies both languages.) Second, substitution routines replace the source language

[4]The approach is "universal" only to the extent that the linguistic theory is "universal." There are some residual phenomena not covered by the theory that are consequently not handled by the system in a principle-based manner. For example, the language-specific English rules of *it-insertion* and *do-insertion* cannot be accounted for by parameterized principles, but must be individually stipulated as idiosyncratic rules of English. Happily, there appear to be only a few such rules per language since the principle-based approach factors out most of the commonalities across languages.

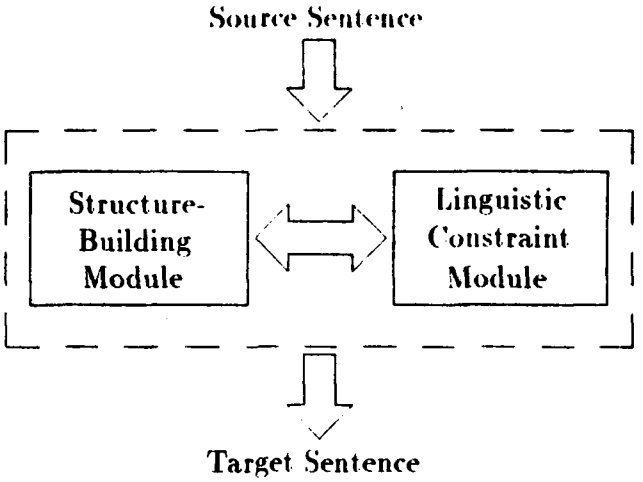


Figure 6: The two top-level modules of the UNITRAN system are the structure-building module and the linguistic constraint module. The structure-building module applies traditional syntactic actions, while the linguistic constraint module enforces well-formedness conditions on the structures passed to it.

constituents with the thematically corresponding target language lexical entries. Third, the generator performs movement and morphological synthesis, thus deriving the target language sentence. An overview of the translation system is given in the next section.

4 Overview of UNITRAN

All three translation stages operate in a co-routine fashion: the flow of control is passed back and forth between a structure-building module and a linguistic constraint module. (See figure 6.) At each of the three stages of translation processing tasks are divided between the two modules as shown in figure 7.

During the parsing stage, the structure-building component, an implementation of the Earley algorithm (1970), applies predicting, scanning, and completing actions, while the linguistic constraint component, an implementation of GB principles, enforces well-formedness conditions on the structures passed to it. The phrase-structures that are built by the structure-building component are underspecified, (*i.e.*, they do not include information about agreement, abstract case, semantic roles, argument structure, *etc.*); the basis of these structures is a set of templates derived during a precompilation phase according to certain source language parameters.[5] The linguistic constraint component eliminates or modifies the underspecified phrase-structures according to principles of GB (*e.g.*, agreement filters, case filters,

[5]The precompilation phase is discussed in Dorr (1987), but is not the focus of this report. In a nutshell, it consists of compiling the principles of a GB subtheory ($\overline{X}$-Theory) concerning phrase structure templates. These templates are generated according to certain parameter settings (*e.g.*, constituent order, choice of specifiers, *etc.*) of the source language. The precompiled phrase structures are then used to drive the parsing mechanism.

Translation Stage	*Structure-Building Tasks*	*Linguistic Constraint Tasks*
Parser	Syntactic Parse: Predict, Scan, Complete	Phrase Structure Constraints: Agreement and Case Filters, Argument Structure and Semantic Role Checking
Substitution	Lexical Replacement	Lexical Constraints: Argument Structure and Thematic Divergence Tests
Generator	Structural Movement and Morphological Synthesis	Structural and Morphological Constraints

Figure 7: The translation tasks of the structure-building and linguistic constraint modules differ according to the stage of the translation. During parsing, the structure-building module performs a syntactic analysis of the source language sentence, while the linguistic constraint module applies structural filters and checks well-formedness. During substitution, lexical replacement is performed by the structure-building module, and tests are applied to predicate-argument structure by the linguistic constraint module. In the generation stage, the structure-building module performs a syntactic synthesis of the target language sentence, while the linguistic constraint module applies structural filters and checks morphological well-formedness.

argument requirements,[6] semantic role conditions, etc.). This design is consistent with several studies that indicate that the human language processor initially assigns a (possibly ambiguous or underspecified) structural analysis to a sentence, leaving lexical and semantic decisions for subsequent processing.[7]

Because the linguistic constraints are available during parsing, the structures built by the structure-building module need not be elaborate; consequently the grammar size need not, and should not, be as large as is found in many other parsing systems. In fact, the number of phrase structure templates that are generated per language generally does not exceed 150 since there are a limited number of configurations per language that are allowed by the principles of $\overline{X}$-Theory. The reduction in grammar size means that the system is not subject to the same slow-downs that are found in other systems. As noted in Barton (1984), in a typical parsing system the description of a language is lengthy, thus increasing the running time of many parsing algorithms. For example, the Earley algorithm (1970) for context-free language parsing can quadruple its running time when the grammar size is doubled. Because the approach here does not employ a lengthy language description, the computational cost of searching the grammar is reduced.

Just prior to the lexical substitution stage, the source language sentence is in an *underlying form*, *i.e.*, a form that can be translated into any target language according to conditions relevant to that target language. This means that all participants of the main action (*e.g.*, *agent*, *patient*, *etc.*) of the sentence are identified and placed in a "base" position relative to the main verb. At the level of lexical substitution, the structure-building module simply

[6] In general, an *argument* of a verb is a subject or an object of the verb, as specified in the verb's dictionary entry.

[7] Frazier 1986 provides recent psycholinguistic evidence that parsing proceeds in this fashion.

replaces target language words with their equivalent target language translations, subject to argument structure requirements and tests of thematic divergence (*i.e.*, tests for semantic mismatches as in the *gustar-like* example mentioned in section 1).

Generation consists of transformation of the sentence into a grammatically acceptable form with respect to the target language (*e.g.*, in English the underlying form *was called John* would be transformed into the surface form *John was called*). An example of how the translator operates is illustrated in the next section.

5 An Example

This section demonstrates the parsing, substitution, and generation stages for translation of the following sentence:

(1) Comió una manzana.
'{He, she} ate an apple.'

5.1 Parsing Stage

As mentioned in section 3.2, there is a "null subject" parameter that is set to TRUE for Spanish. The parser must access this parameter to "know" that a missing subject in (1) does *not* rule out the sentence (as it would in English). Figure 8 gives snapshots of the parser in action. First the Earley structure-building component predicts that the sentence has a noun phrase (NP) and a verb phrase (VP) (see (a)), the order of which is determined by the "constituent order" parameter at precompilation time.[8] The only structures available for prediction by the Earley module are those generated at precompilation time; thus, at this point no further information about the structure is available until the linguistic constraint module takes control.

The constraint module accesses the "null subject" parameter, which dictates that the empty element attached to NP is a subject; the [+pro] (pronominal) feature is associated with the node (see (b)) so that this position will accommodate both null subject source languages and overt subject source languages.[9]

In snapshot (c), the Earley module expands VP and scans the first input word *comer*. Now the Earley module cannot proceed any further; thus, the constraint module takes over again. First a semantic role (or θ-role, as it is called in GB Theory) of *agent* is assigned to the empty subject of the sentence. This information is determined from the dictionary entry of *comer* which dictates that this verb requires both an agent (assigned to the subject or *external argument* of the verb) and a theme (assigned to the object or *internal argument* of the verb). The dictionary entry for *comer* is encoded as follows:

```
(comer: [ext: agent] [int: theme] V (english: eat) (french: manger) ...)
```

[8]Since Spanish is a *head-initial* language, NP must precede VP. This would not be the case for non-*head-initial* languages.

[9]For example, Italian and Hebrew do not require an overt subject, but English and French do; thus, during a later stage (generation), e[pro] will either be left as is, or lexicalized to a pronominal form (*e.g.*, *he* or *she* in English) that agrees with the main verb.

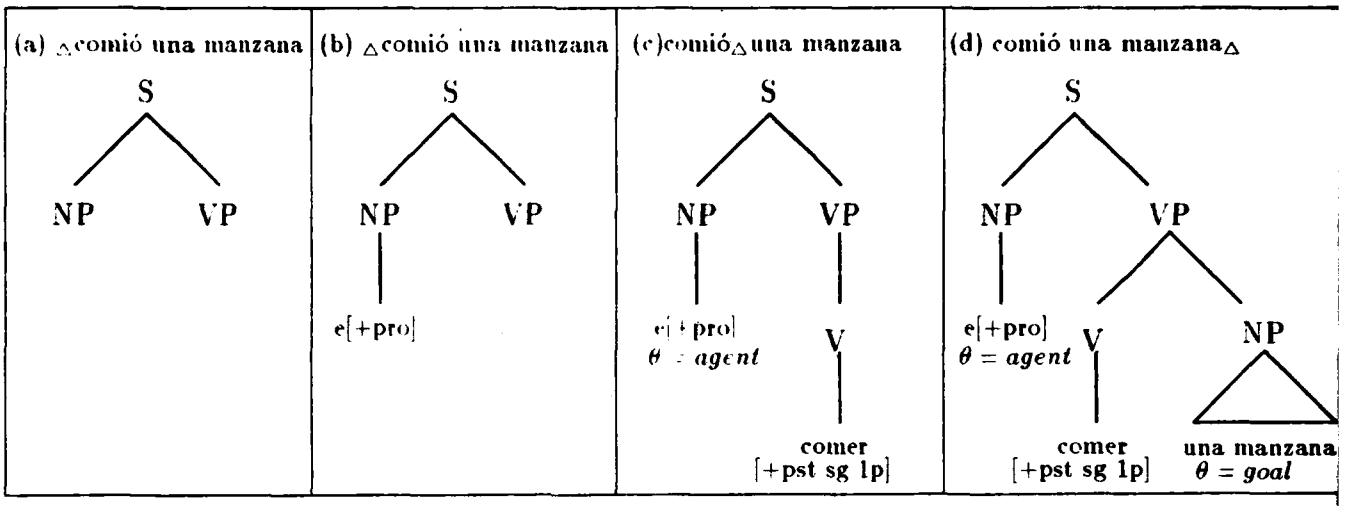
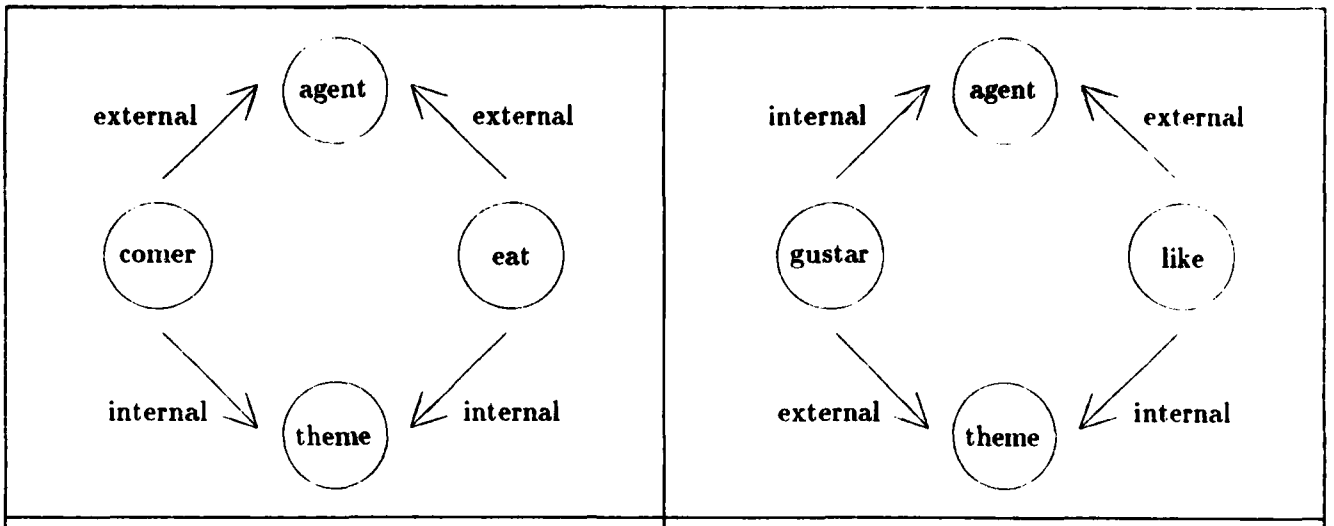


Figure 8: First, Earley predicts that the sentence has an NP and a VP (see (a)). Next, the "null subject" parameter dictates that the empty element under NP is a subject (see (b)). At this point Earley expands VP, thus scanning the first input word and assigning *agent* θ-role (see (c)). Finally, the internal argument NP is parsed and the *goal* θ-role is assigned (see (d)). In the resulting underlying (interlingual) form all participants (*agent* and *theme*) of the main action (*comer*) have been identified, and all arguments (subject and object) are in their "base" positions (external and internal) with respect to the verb *comer*. The verb *comió* has been changed to the infinitive form *comer* (with person, tense, and number features) via a morphological analysis stage that will not be discussed here.

In order to parse the final two words, the constraint module first predicts that a noun phrase (corresponding to the internal argument of *comer*) follows the verb. Then the Earley module scans the final two words, thus completing the NP and allowing the constraint module to assign a θ-role of *theme* to *una manzana*. Snapshot (d) shows the completed parse. The sentence is now in the underlying (interlingual) form required for the substitution and generation phases. That is, all participants (*agent* and *theme*) of the main action (*comer*) have been identified, and all arguments (subject and object) are in their "base" positions (external and internal) with respect to the verb *comer*. The equivalent target language sentence can now be derived via the synthesizer (which is programmed to operate on the basis of the target language parameter settings).

5.2 Substitution Stage

There are two parts to the substitution stage. First, a mapping between thematic roles takes place. That is, the argument structure of the source language verb *comer* is examined to determine the position of the *agent* and the *theme* for the target language verb *eat*. In the example presented here, the positioning of *agent* and *theme* are the same for both Spanish and English, *i.e.*, the agent is external and the theme is internal in both cases. Thus, the thematic divergence test is not required; the agent and theme are directly translated *in*



comer: Yo como el pan. **eat:** I eat the bread.	**gustar:** El libro me gusta a mí. **like:** I like the book.

Figure 9: There is no thematic divergence between the Spanish verb *comer* and its English equivalent *eat* since the internal and external arguments match in θ-role. By contrast, the Spanish verb *gustar* and its English equivalent *like* exhibit thematic divergence since the θ-roles of the internal and external arguments are reversed.

situ. However, this direct mapping does not always apply, *e.g.*, in the case of the *gustar-like* divergence discussed in section 1. Figure 9 illustrates the distinction between the argument structures of *comer* and *gustar*. In such cases of thematic divergence, a more complex mapping is required.

The second part of the substitution stage is lexical replacement. All arguments and actions are replaced by the corresponding equivalent forms found in the lexical entries of the source language words. The structure resulting from substitution is shown in figure 10.

5.3 Generation Stage

Generation is both structural and morphological. First, structural routines check to see whether movement (*e.g.*; passivization, raising, *etc.*) is required. Because the sentence is a simple active sentence, no such movement is required. Next, morphological routines take over to generate the correct form of the main verb, and also to realize the subject of the sentence, which up until this point has been empty. In order for this realization (or *lexicalization*) to take place, the generator must "know" that English requires a subject — otherwise, the subject will be left incorrectly unrealized. Thus, the "null subject" parameter mentioned in section 3.2 is accessed at generation time. The final target language sentences are:

(2) He ate an apple.
She ate an apple.

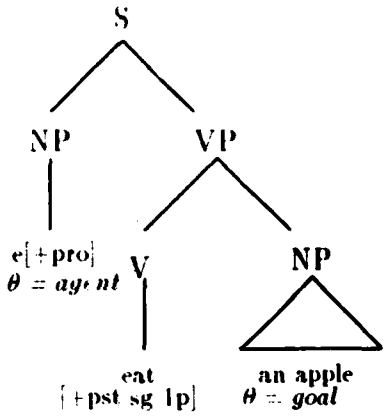


Figure 10: After the substitution stage, the underlying form has been lexically modified to accommodate the target language. First a mapping between thematic roles identifies the *agent* and *theme*. Next lexical replacement of target language words for source language words establishes that the main verb is *eat* and its internal argument is *an apple*.

Note that the form e[+pro] has been lexicalized as both *he* and *she* to match the person and number of the verb eat. The translation has revealed an ambiguity that exists implicitly in the Spanish source sentence: without context, the subject of the Spanish sentence may be interpreted as either *he* or *she*.

6 Conclusions and Future Work

The system described here is based on modular theories of syntax which include systems of principles and parameters rather than complex, language-specific rules. The contribution put forth by this investigation is two-fold: (a) from a linguistic point of view, the investigation allows the principles of GB to be realized and verified; and (b) from a computational perspective, descriptions of natural grammars are simplified, thus easing the programmer's and grammar writer's task. The model not only permits a language to be described by the same set of parameters that specify the language in linguistic theory, but it also eases the burden of the programmer by handling interaction effects of universal principles without requiring that the effects be specifically spelled out.

Currently the UNITRAN system operates bidirectionally between Spanish and English; other languages may easily be added simply by setting the parameters to accommodate those languages.[10] The system operates with a success rate of approximately 80 per cent. The time to translate an average length sentence is approximately 30-50 seconds, depending on the complexity of the phenomena encountered.

Experiments are underway to determine the "optimal" balance of principle clustering between the precompilation and processing phases. The question under investigation is how much structure must be generated at precompilation time in order to efficiently perform on-line verification of GB constraints. On the one hand, incorporating a large number of

[10] Experiments with Warlpiri and other "non-standard" languages are currently underway.

constraints into the precompilation phase causes the grammar size to become explosive, thus slowing down grammar search time; on the other hand, eliminating a large number of constraints from precompilation forces a high cost at constraint verification time. In the present incarnation of the parser presented here, a relatively small number of GB constraints (those concerning skeletal phrase structures and empty noun phrases) are accessed at precompilation time, leaving many of the GB constraints to apply at processing time. Timing tests have shown this clustering of principles to be promising for the interlingual design presented here. Ultimately, the goal is for a small set of principles (grouped into modules) to cover phenomena found in all languages so that unmanageable grammar size is no longer a problem.

7 References

Barton, Edward G. Jr. (1984) "Toward a Principle-Based Parser," MIT AI Memo 788.

Chomsky, Noam A. (1981) *Lectures on Government and Binding*, Foris Publications, Dordrecht.

Dorr, Bonnie J. (1987) "UNITRAN: A Principle-Based Approach to Machine Translation," AI Technical Report 1000, Master of Science thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Earley, Jay (1970) "An Efficient Context-Free Parsing Algorithm," *Communications of the ACM* 14, 453-460.

Frazier, Lyn (1986) "Natural Classes in Language Processing," presented at the *Cognitive Science Seminar, MIT, November*, Cambridge, MA.

Gazdar, G., E. Klein, G. Pullum, and I. Sag (1985) *Generalized Phrase Structure Grammar*, Basil Blackwell, Oxford, England.

Sharp, Randall M. (1985) "A Model of Grammar Based on Principles of Government and Binding," M.S. thesis, Department of Computer Science, University of British Columbia.

Slocum, Jonathan (1984) "METAL: The LRC Machine Translation System," presented at the *ISSCO Tutorial on Machine Translation, Lugano, Switzerland*, Linguistics Research Center, University of Texas, Austin.

Slocum, Jonathan and Winfield S. Bennett (1985) "The LRC Machine Translation System," *Computational Linguistics* 11:2-3, 111-121.

van Riemsdijk, Henk and Edwin Williams (1986) *Introduction to the Theory of Grammar*, MIT Press, Cambridge, MA.

END
DATE
FILMED
7-88
DTIC