UNITRAN: An Interlingual Approach to Machine Translation

Bonnie Dorr

M.I.T. Artificial Intelligence Laboratory

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 in part because interaction effects of complex phenomena made translation appear to be unmanageable. Later approaches to the problem have succeeded but are based on many language-specific rules. To capture all natural language phenomena, rule-based systems require an overwhelming number of rules; thus, such translation systems either have limited coverage, or poor performance due to formidable grammar size. This paper presents an implementation of an "interlingual" approach to natural language translation. The UNITRAN system relies on principle-based descriptions of grammar rather than rule-oriented descriptions.[2] The model is based on linguistically motivated *principles* and their associated *parameters* of variation. Because a few principles cover all languages, the unmanageable grammar size of alternative approaches is no longer a problem.

I. Introduction

The problem addressed in this paper is to construct a translation model that operates cross-linguistically without relying on complex language-specific rules. Many machine translation systems depend heavily on context-free rule-based systems. For example, the METAL system [Slocum, 1984], [Slocum and Bennett, 1985] is a transfer approach that relies on a large database of rules per language, solely for syntactic processing. The aim of this paper 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 from 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 other designs will be discussed. Finally, an overview of the UNITRAN design will be given, and a translation example will be shown.

Verb Preposing:
¿Qué vio Juan?
'What did John see?'
Null Subject:
Vio al hombre.
'{He, She} saw the man.'[3]
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 comió la cena salió.'

Table 1: Sentences handled by UNITRAN

The model that has been constructed is based on abstract principles of the "Government and Binding" (GB) [Chomsky, 1981] framework. The grammar is viewed as a modular system of principles rather than a large set of language-specific rules. Distinctions among languages are handled by settings of parameters associated with the principles. Several types of phenomena are handled without sacrificing cross-linguistic application (table 1 shows some examples).

The system gives the user access to parameter settings, thus enabling additional languages to be handled. Interaction effects of the principles are handled by the system, not the user, thus eliminating the task of spelling out the details of rule applications. Before the source language processing (parsing) takes place, the parameters are set according to the source language values, and are then *reset* according to the target language values before target language processing (generation) occurs. For example, a "constituent order" parameter is associated with a universal principle that requires a language-dependent ordering of constituents with respect to a phrase. The user should

[1]This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for this work 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.

[2]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.

[3]The "{.. , ..}" notation denotes optionality. Thus, the subject of the sentence may either be *he* or *she*.

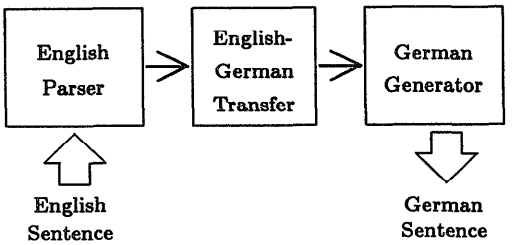

Figure 1: Transfer Translation Approach as found in METAL (1984)

set this parameter to be *head-initial* for a language like English, but *head-final* for a language like Japanese.

Translation is primarily syntactic; thus, there is no global contextual "understanding" (the system translates one sentence at a time). 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 arguments of verbs.[4] It should be noted that determining the mapping between arguments of semantically equivalent verbs is nontrivial.[5] For example, although the Spanish verb *gustar* is semantically equivalent to the English verb *like*, the argument structures of these two verbs differ. The subject of *like* is the *agent*, whereas the object of *gustar* is the *agent*. Because of 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.

II. Transfer vs. Interlingual

This section compares a non-interlingual (rule-based) system to the interlingual (principle-based) design of UNITRAN.

A. A Transfer Approach: METAL

The *transfer* approach to translation has been taken in METAL [Slocum, 1984], [Slocum and Bennett, 1985]. In this approach there is a parser and a generator for each source and target language. In addition, there are a set of *transfer* components, one for each source-target language pair (see figure 1). The transfer phase is actually a third translation stage in which one language-specific representation is mapped into another. The METAL system currently translates from German into Chinese and Spanish, as well as from English into German.

[4]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.

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

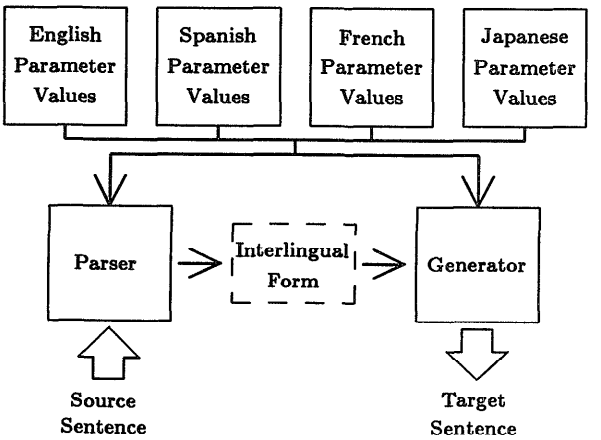

Figure 2: The Interlingual Design of UNITRAN

The malady of the transfer approach is that each of the parsing, generation and transfer components is entirely language-specific.[6] Because the system has no access to universal principles, there is no consistency across the components; thus, each component 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 component 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 parse time. Furthermore, there is no "rule-sharing" — all rules apply to only one language.

B. An Interlingual Approach: UNITRAN

The translation model described in this paper 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 rules.

The interlingual approach has been taken in the past [Sharp, 1985]; however, the UNITRAN system differs from

[6]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 parser is nevertheless based on a large database of language-specific rules. For example, the German parser is based on phrase-structure grammar, augmented by procedures for transformations, and the English parser employs a modified GPSG approach.

Parameter	*Spanish*	*English*
Constituent Order	head-initial	head-initial
Null Subject	TRUE	FALSE
Clitic Doubling	TRUE	FALSE
Inversion	V-preposing	Subject-Aux

Table 2: Parameter Values for Spanish and English

Constituent Order:
The man ate cheese.
*The man cheese ate.[7]
Null Subject:
Vio al hombre.
*Saw the man.
Clitic Doubling:
Juan lo vio al hombre.
*John him saw the man.
Inversion:
*What saw John.
Qué vio Juan.
What did John see.
*Qué hizo Juan ver.

Table 3: Effects of Parameter Settings for Spanish and English

Sharp's system in three respects. First, the system uses the same parser and generator for all languages, whereas Sharp's system requires the user to supply parser for each source language and a generator for each target language. Second, the user is allowed to specify parameter values to the principles — thus modifying the effect of the principles from language to language — while in Sharp's system, the user has limited access to the parameters of the system (*e.g.*, the "constituent order" parameter mentioned in section I is not available for modification). Third, the system generates rules on the fly using linguistically motivated principles; by contrast, in Sharp's system context-free rules (set up for English-like languages) are hardwired into the code; 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 approach presented here 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.[8] Figure 2 illustrates the design of the model.

[7]The equivalent structure for **the man cheese ate* (= **el hombre queso comió*) is illegal in Spanish also. On the other hand, the sentence is legal for Japanese and other head-final languages.

[8]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.

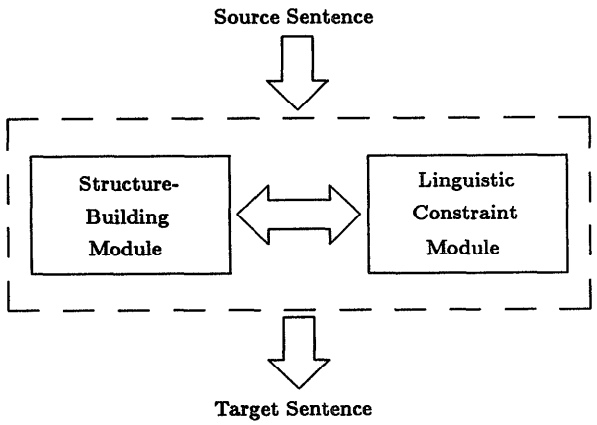

Figure 3: Structure-Building and Linguistic Constraint Modules of UNITRAN

The parser and generator are user-programmable: all of the principles associated with the system are associated with parameters that are set by the user. Thus, the user does not need to supply a source language parser or a target language generator since these are already part of the translation system. The only requirement is that the built-in parser and generator be *programmed* (via parameter settings) to process the source and target languages. For example, the user 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; by contrast, this parameter must be set to FALSE for English. (For details on the null subject parameter, see [van Riemsdijk and Williams, 1986].) Table 2 shows some examples of the parameters and their settings for Spanish and English. Table 3 describes the effects of each these parameters respectively.[9] A dictionary for each language must also be supplied. The next section describes the system in more detail.

III. Overview of UNITRAN

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 constituents with the thematically corresponding target language lexical entries. Third, the generator performs movement and morphological synthesis, thus deriving the target language sentence.

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 con-

[9]An asterisk (*) denotes ill-formedness.

Stage	*Tasks*
Parsing	**Structure-Building Tasks:** Predict, Scan, Complete
	Linguistic Constraint Tasks: Agreement and Case Filters, Argument Structure and Semantic Role Checking
Substitution	**Structure-Building Tasks:** Lexical Replacement
	Linguistic Constraint Tasks: Thematic Divergence Tests and Argument Structure Completion
Generation	**Structure-Building Tasks:** Structural Movement and Morphological Synthesis
	Linguistic Constraint Tasks: Structural and Morphological Well-formedness Tests

Table 4: Translation Tasks of Structure-Building and Linguistic Constraint Modules

straint module. (See figure 3.) At each of the three stages of translation, processing tasks are divided between the two modules as shown in table 4.

During the parsing stage the structure-building component, an implementation of the Earley algorithm (see [Earley, 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.[10] The linguistic constraint component eliminates or modifies the underspecified phrase-structures according to principles of GB (*e.g.*, agreement filters, case filters, argument requirements, 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. (See [Frazier, 1986].) 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.[11] Thus, the system avoids computational costs due to large grammar size.

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 replaces target language words with their equivalent target language translations while the linguistic constraint module applies tests for for semantic mismatches as in the *gustar-like* example mentioned in section I, and fulfills argument structure requirements.

During generation, the structure-building module transforms the sentence into a grammatically acceptable form with respect to the target language; in English the underlying form *was called John* would be transformed into the surface form *John was called*. Tests for grammaticality are made by the linguistic constraint module according to structural and morphological constraints, which are parameterized to satisfy the target language requirements.

IV. 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.'

A. Parsing Stage

As mentioned in section II.B, 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 4 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.[12] 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

[10]The precompilation phase is discussed in [Dorr, 1987], but is not the focus of this paper. 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.

[11]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. Thus, the running time of the parser 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, Earley algorithm for context-free language parsing can quadruple its running time when the grammar size is doubled.)

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

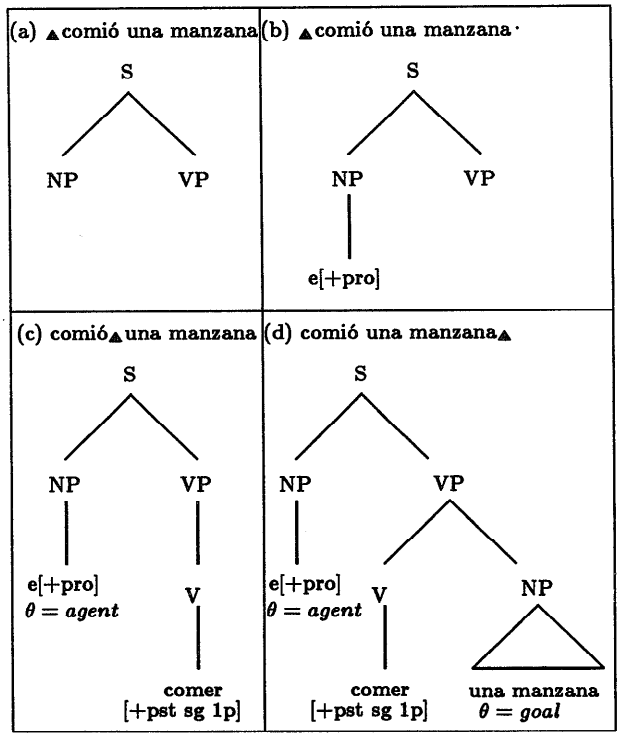

Figure 4: Snapshots of Parser in Action

to NP is a subject; the [+pro] (pronominal) feature is associated with the node (see (b)) so that subject will accommodate both null subject source languages and overt subject source languages.[13]

In snapshot (c), the Earley module expands VP and scans the first input word *comer*.[14] 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) ...)
```

[13]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.

[14]The verb *comió* has been changed to the infinitive form *comer* (with person, tense, and number features) via a morphological analysis that precedes the parsing stage. The details of the morphological analysis stage will not be discussed here.

Lexical Entry	*Sample Usage*
comer: [ext: agent] [int: theme] ...	Yo como el pan.
eat: [ext: agent] [int: theme] ...	I eat bread.
gustar: [ext: theme] [int: agent] ...	El libro me gusta a mí.
like: [ext: agent] [int: theme] ...	I like the book.

Table 5: Thematic Correspondence (Comer and Eat) vs. Thematic Divergence (Gustar and Like)

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 source language sentence can now be derived via the generator (which is programmed to operate on the basis of the target language parameter settings).

B. 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 situ*. However, this direct mapping does not always apply, *e.g.*, in the case of the *gustar-like* divergence discussed in section I.

Table 5 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 verbs and arguments are replaced by the corresponding equivalent forms found in the lexical entries of the source language words. The resulting target language underlying form is shown in figure 5.

C. 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

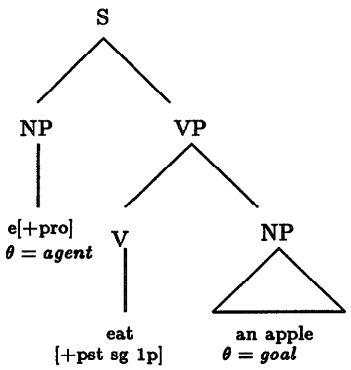

Figure 5: Target Language Underlying Form

English requires a subject — otherwise, the subject will incorrectly be left unrealized. Thus, the "null subject" parameter mentioned in section II.B is accessed at generation time. The final target language sentences are:

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

V. Conclusions

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.[15]

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

Acknowledgements

I would like to thank Bob Berwick, Ed Barton, Sandiway Fong and Dave Braunegg, all of whom provided useful guidance and commentary during this research.

References

[Barton, 1984] Barton, G. Edward, Jr. *Toward a Principle-based Parser.* Technical Report AI Memo 788, Massachusetts Institute of Technology, July 1984.

[Chomsky, 1981] Noam A. Chomsky. *Lectures on Government and Binding, the Pisa Lectures.* Volume 9 of *Studies in Generative Grammar*, Foris Publications, Dordrecht, 1981.

[Dorr, 1987] Bonnie J. Dorr. *UNITRAN: A Principle-Based Approach to Machine Translation.* Master's thesis, Massachusetts Institute of Technology, 1987.

[Earley, 1970] Jay Earley. "An Efficient Context-Free Parsing Algorithm." *Communications of the ACM*, 13, 1970.

[Frazier, 1986] Lyn Frazier. Natural Classes in Language Processing. November 1986. presented at Cognitive Science Seminar, MIT.

[Sharp, 1985] Randall M. Sharp. *A Model of Grammar Based of Principles of Government and Binding.* Master's thesis, The University of British Columbia, October 1985.

[Slocum, 1984] Jonathan Slocum. "METAL: The LRC Machine Translation System." In *Proceedings of ISSCO Tutorial on Machine Translation*, Lugano, Switzerland, 1984.

[Slocum and Bennett, 1985] Jonathan Slocum and Winfield S. Bennett. "The LRC Machine Translation System." *Computational Linguistics*, 11:111–121, 1985.

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