REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

REPORT NUMBER: AIM-947	2 GOVT ACCESSION NO
3 RECIPIENT'S CATALOG NUMBER	
TITLE (and Subtitle): Principle-Based Parsing For Machine Translation	5 TYPE OF REPORT & PERIOD COVERED: AI Memo 9/84-5/87
6 PERFORMING ORG REPORT NUMBER	
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: 17 (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 translation; parsing; principles vs. rules; co-routine design; linguistic constraints.

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

See back.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Number 20.

This report shows how a principle-based parser with a "co-routine" design improves parsing for translation. The parser consists of a skeletal structure-building mechanism that operates in conjunction with a linguistically based constraint module, passing control back and forth until a set of underspecified skeletal phrase-structures is converted into a fully instantiated parse tree. The modularity of the parsing design accommodates linguistic generalization, reduces the grammar size, allows extension to other languages, and is compatible with studies of human language processing.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 947 December, 1987

PRINCIPLE-BASED PARSING FOR MACHINE TRANSLATION

Bonnie J. Dorr

ABSTRACT:

Many syntactic parsing strategies for machine translation systems are based entirely on context-free grammars. These parsers require an overwhelming number of rules; thus, translation systems using rule-based parsers either have limited linguistic coverage, or they have poor performance due to formidable grammar size. This report shows how a principle-based parser with a "co-routine" design improves parsing for translation. The parser consists of a skeletal structure-building mechanism that operates in conjunction with a linguistically based constraint module, passing control back and forth until a set of underspecified skeletal phrase-structures is converted into a fully instantiated parse tree. The modularity of the parsing design accommodates linguistic generalization, reduces the grammar size, allows extension to other languages, and is compatible with studies of human language processing.

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, Dave Braunegg, Bruce Dawson, and Sandiway Fong. This report is an extended version of a paper that is in the Proceedings of the Ninth Annual Conference of the Cognitive Science Society (1987).

1 Introduction

This report explains how to construct a syntactic parsing model that accommodates cross-linguistic uniform machine translation without relying on language-specific context-free rules. Parsing systems typically use grammars that describe language with complicated rules that spell out the details of their application. ATN-based systems (Woods, 1970; Bates, 1978) have several hundred grammar arcs, each with detailed tests and actions; augmented phrase-structure grammars as used in Diagram (Robinson, 1982) spell out the type, position, and probability of occurrence of constituents in a given phrase; and the GPSG approach (Gazdar *et. al.*, 1985) uses a "slash-category" mechanism to incorporate long-distance relations directly into the grammar rules.[1] Such systems do not work in the context of translation across several languages: the rules of a given grammar are painstakingly tailored to describe a *single* language, thus forcing a loss of linguistic generalization and limiting the addition of new languages.[2]

An additional problem with rule-based systems is that the grammar size is typically quite formidable. For example, Slocum's METAL system (1984, 1985), developed at the Linguistics Research Center at the University of Texas, relies on thousands of context-free rules per language solely for parsing. Each parser operates unilingually and accesses an unwieldy number of language-specific rules. Unfortunately, the grammar size of a parsing system makes a difference in processing time. As noted in Barton (1984), the Earley algorithm (1970) for context-free language parsing can quadruple its running time when the grammar size is doubled.

Another disadvantage of rule-based systems is that they fail to preserve the modular organization of new theories of grammar. Designing a system on the basis of a rule-based linguistic theory means that the grammar writer must keep track of hundreds of rules and the context in which each rule applies in order to do any system editing. Preserving modularity allows general conditions to be factored out so that each system component is simplified and language descriptions are reduced in size. Furthermore, modularity allows several people to work on the same system without affecting one another, since each is working on an independent component of the system.

In this report I describe an implementation of a parsing model that is based on subsystems of grammatical principles and parameters.[3] The parser follows a "co-routine" design: the structure-building mechanism operates with access to linguistic constraints of Govern-

[1] Barton (1984) describes these rule-based systems in more detail.

[2] GPSG *does* make use of constraints that are claimed to be cross-linguistically applicable (see Gazdar *et. al.*, 1985, p. 4.). However, the universals used by GPSG (for example, the Exhaustive Constant Partial Ordering constraint on linear precedence in grammar) follow as a consequence of the grammatical formalism itself; they do not necessarily follow from empirical data. Thus, the constraints of GPSG differ from GB in that they are not developed on the basis of observation of natural language phenomena, but they are derived from formal statements of the grammatical metalanguage. Furthermore, the cross-linguistic applicability of GPSG is not as readily observable as that of GB since there is no notion of parameterized linguistic principles; instead, there are many complex and idiosyncratic grammar rules that are difficult to decode without understanding the intent of the grammar-writer.

[3] 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. The parameter is set by the grammar-writer to be *head-initial* for a language like English, but *head-final* for a language like Japanese. This is discussed in section 2.1.

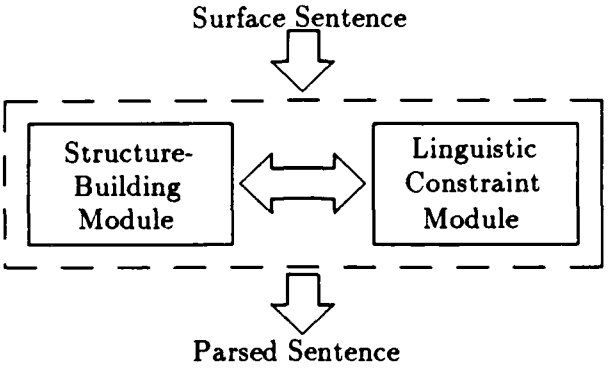

Figure 1: The parser takes on a co-routine design. The structure-building module constructs skeletal syntactic structures; these are then modified by the linguistic constraint module according to the principle of GB. The two modules pass control back and forth until the sentence is completely parsed.

ment and Binding (GB) theory as developed by Chomsky (1981, 1982). (See figure 1.) The structure-building module assigns a skeletal syntactic structure to a sentence, and then this structure is eliminated or modified according to the principles of GB. This design is consistent with recent psycholinguistic studies that indicate that the human processor initially assigns a (potentially ambiguous or underspecified) structural analysis to a sentence, leaving semantic descriptions for subsequent processing.[4] Furthermore, the parser is designed so that it applies uniformly across many languages, allowing the grammar-writer to modify the parameters of the system to accommodate additional languages. Currently, the system operates bidirectionally between English and Spanish.

Parsing uniformly across languages is difficult because the parser appears to require a massive amount of "knowledge." Not only must it be able to parse several types of phenomena (and their interaction effects) in a language, but it must also avoid giving ill-formed sentences the same status as well-formed sentences.[5] Consider (1):

(1) Le quiere a Juan
'(She) loves John'

Although (1) appears to be simple, it is not simple from the viewpoint of uniform parsing since the equivalent sentence parses differently in other languages. The Spanish and English parse trees for (1) are in figure 2. Literally, the English translation for (1) is (2), which is ungrammatical:

(2) him *e* loves to John

[4]Frazier (1986) provides recent psycholinguistic evidence that there is a temporal sequence of parsing consistent with the GB-based model presented here. However, the issue of psycholinguistic reality of the model is not the central focus of this report.

[5]Partial sentences are ignored here. A system that performs question-answering allows partial sentences to be parsed as well-formed structures. The system described here analyzes sentences in isolation. Thus, incomplete sentences are considered ill-formed.

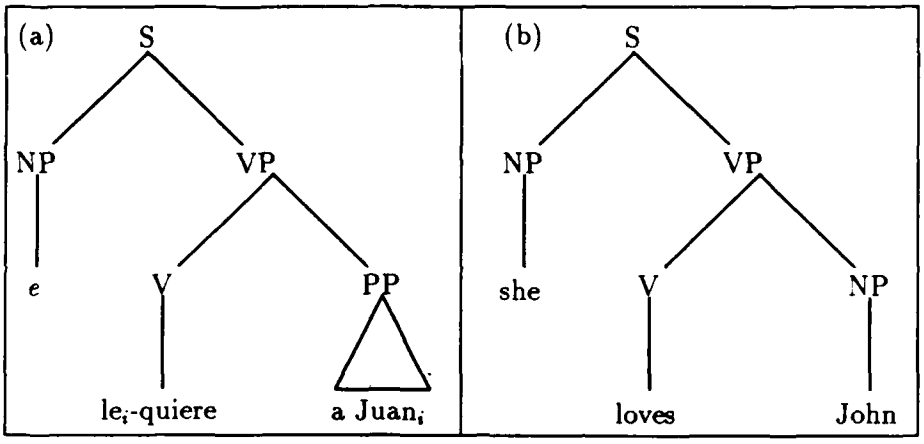

Figure 2: The Spanish and English parse trees for equivalent sentences are not always identical. For example, here the subject is not lexically realized in the Spanish parse tree, but it is overt in the English parse tree. (Subscripts are used for co-referring elements; thus, *le* (= him) refers to *Juan*.)

The *e* stands for a null subject that is realized as *she* in English. (See section 2.3 for a discussion of the null subject phenomenon in Spanish.) The parsing implementation presented here rules out sentence (2) without sacrificing the ability to parse (1).

Perhaps a more important consideration than ruling out ungrammatical sentences is the requirement that the parser avoid assigning wrong interpretations to grammatical sentences. In a cross-linguistically applicable system, this requirement is difficult to satisfy. For example, it is conceivable that the system might parse a Spanish sentence incorrectly on the basis of the knowledge it has for parsing English sentences. Consider (3):

(3) Qué golpeó Juan
'What did John hit'

If the parser were to use English parameter settings to parse this sentence, it would understand the sentence to mean *what hit john* (*i.e.*, the *agent* and *goal* roles would be reversed). The parameter-setting approach allows incorrect interpretations such as this one to be avoided in one language without affecting the processing of other languages.

The co-routine design differs from other GB parsing/translation systems (*e.g.*, Sharp, 1985) in that the linguistic principles are used for "on line" verification during parsing rather than as well-formedness conditions on output. Furthermore, in Sharp's system, context-free rules (set up for English-like languages) are hardwired into the code rather than generated by the parser from principles of GB; thus, Sharp's system cannot handle languages (like German or Japanese) that do not have the same order of constituents as English. This malady comes about in Sharp's system because the grammar-writer has limited access to the grammatical principles of the system. The system described here allows the grammar-writer to specify parameter values to the principles, thus modifying their effects from language to language. Some GB principles are applied on line (*i.e.*, at processing time), while others are applied off

line (*i.e.*, at precompilation time).[6] Both classes of principles include parameters of variation.

The modularity imposed by the GB framework is an improvement over context-free based systems for several reasons. First, properties common to several languages are not specified directly in rules, but are abstracted into modularized principles. For example, the passive transformation that relates an active sentence to its passive counterpart used to look something like the following:

(4) NP_1 V $NP_2 \Rightarrow NP_2$ be V+en by NP_1

Thus, the sentence *Susan beat John* is related by the passive transformation to *John was beaten by Susan*. Rule (4) is complicated and idiosyncratic. It relies heavily on the word choice and ordering requirements of English. Unfortunately, word choice and ordering do not necessarily carry over to other languages. In Spanish, there are three passive transformations:

(5) NP_1 V $NP_2 \Rightarrow NP_2$ ser V+ido por NP_1
NP_1 V $NP_2 \Rightarrow$ se V a NP_2
NP_1 V $NP_2 \Rightarrow$ se le/les V

Only the first of the three Spanish passive transformations in (5) is the same as the one English passive transformation. Thus, the sentence *John was beaten by Susan* can be literally translated as *John fue golpeado por Susan*. However, the passive form may also be realized as *se golpeó a Juan* (here the subject is not specified) or *se le golpeó* (here the subject and object are not specified).

The abstraction of properties into modularized principles allows linguistic generalization to be captured. The system uses a general principle called move-α rather than a detailed passive rule that changes from language to language. This movement principle allows a constituent (α) to be displaced to another position in the sentence, but the movement is constrained according to principles of Trace Theory (to be discussed in section 2.3). Because these constraining principles are allowed to vary from language to language, we can account for the fact that Spanish passive NP-movement may involve realization of a pronoun *se*, whereas the English passive NP-movement does not allow such a realization. Thus, the passive rule is reduced to a small set of cross-linguistically applicable principles that are sensitive to parametric variation.

Another advantage to modularity is that multiplicative effects of linguistic constraints are not spelled out in the form of grammar rules. In a rule-based system, subject/verb agreement might use the following two rules:

[6] Experiments are currently underway to determine the "optimal" balance of principle clustering between the precompilation and processing phases. In order for the linguistic constraints to apply, a structure must first be created. The question under investigation is how much structure must be generated at precompilation time in order to perform on line verification of linguistic constraints efficiently. On the one hand, incorporating a large number of 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. Frazier (1986) suggests that all phrase-structure possibilities get multiplied out, leaving only a small subset of GB constraints to apply at processing time. In 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. Time tests have shown this clustering of principles to be promising for the co-routine design presented here.

(6) $S \Rightarrow NP_{sg}\ VP_{sg}$
$S \Rightarrow NP_{pl}\ VP_{pl}$

These two rules work for parsing active sentences, but to also parse passive sentences, subject/verb agreement has to be encoded in passive rules too:

(7) $S \Rightarrow NP_{sg}\ be_{sg}\ VP_{+en}$
$S \Rightarrow NP_{pl}\ be_{pl}\ VP_{+en}$

Now, if another phenomenon (say, past/present tense alternation) is added, each of (6) and (7) will have to be multiplied out into additional rules. It is easy to see that the grammar can quickly become explosive. The more desirable approach is to use a simple (underspecified) grammar, and then superimpose separate modules that individually handle agreement and movement phenomena on the grammar. The elimination of multiplicative effects from the grammar rules allows grammar size (hence processing time) to be reduced.

Modularity has the further advantage that a separate description is not required for each language handled by the system. The grammar-writer does not have the traditional task of constructing a set of complex language-specific phrase-structure rules; instead, the task of the grammar-writer is to determine the parameter-settings for each language. For example, two rules accounting for the fact that a Spanish sentence does not require a subject are the following:

(8) $S \Rightarrow NP\ VP$
$S \Rightarrow VP$

Rather than specifying these two rules, the grammar-writer need only set the *pro-drop* parameter (to be discussed in section 2.3) to T for Spanish. The parameter-setting approach facilitates the extension of the system to handle additional languages: adding a language reduces to changing the parameter-settings to suit that language.

The following sections describe the syntactic parsing model in more detail. Section 2 presents the underlying linguistic theory; section 3 discusses the implementation; section 4 provides an example of the parser in action; and section 5 contains a summary and limitations.

2 Underlying Linguistic Theory

The structure-building and linguistic component of figure 1 correspond to a bi-partition of several underlying subsystems of grammar. The partition corresponding to the structure-building component consists of the $\overline{X}$ subsystem, which imposes certain restrictions on the order and positioning of phrasal constituents. The partition corresponding to the linguistic component consists of the θ and Trace subsystems (as well as others not discussed here), which impose restrictions on movement of constituents in a sentence. The interaction of the subsystems underlying the two components is precisely what is needed to gain the effects of complicated rule systems without stipulatory rules.

This section describes three GB subtheories ($\overline{X}$-Theory, θ-Theory, and Trace Theory) that underlie the two components of the system. The principles and parameters of variation

associated with these three theories are described. Also, the relevance of the parameters within the context of the parsing model is discussed. The goal is to incorporate the parameterized principles of GB into a single, cross-linguistically uniform parsing system.

2.1 $\overline{\text{X}}$-Theory Parameters: Choice of Specifiers and Constituent Order

There are two central notions associated with $\overline{\text{X}}$-Theory. First, the dictionary (henceforth *lexicon*) specifies subcategorization frames for lexical items. For example, the frame for the verb *put* includes two arguments: a noun phrase and a prepositional phrase (*e.g.*, *put the car in the garage*). Second, phrase-structure is expressed as a projection of a lexical head X (= N, V, P or A). Thus, in the sentence *he put the car in the garage*, the verb *put* projects the verb phrase *put the car in the garage*.[7] $\overline{\text{X}}$-Theory assumes that phrase-structures for English are derived by rules of the following form:

(9) $X^{max} \Rightarrow$ (Specifier) X (Complement)

Here X^{max} is the maximal projection of the lexical head X (more commonly called XP). The Specifier of X is determined by a parameter setting associated with the $\overline{\text{X}}$ module, and the complement of X is determined by the subcategorization frame of the verb. For example, if X is a noun, X^{max} is NP, a possible Specifier is a determiner, and a possible complement is a prepositional phrase (depending on whether this is specified in the lexical entry for the noun).

English requires that specifiers of all lexical categories occur before the lexical head and complements follow the lexical head. However, this rule does not apply to all languages (*e.g.*, Navajo, German, Japanese, *etc.*). For example, consider the following Navajo sentence:

(10) ashkii at'ééd yiyiiltsą́
'the boy saw the girl'

This sentence literally translates as *the boy the girl saw* since Navajo requires the complement to precede the head.[8] It is assumed that the constituent order of a language is determined by a parameter of variation. Thus, before parsing begins, $\overline{\text{X}}$ rules are set up according to the constituent order of the language being parsed. This is crucial in the parsing model since many of the principles of other GB subtheories cannot apply until a valid licensed structure (with predetermined ordering restrictions) has first been built. In other words, $\overline{\text{X}}$-Theory provides basic templates to which remaining parsing constraints can apply.

[7]The lexical representation used in the parser presented here is based on the input representation required by the morphological analyzer. It includes the root forms of words and pointers to applicable affixes. Root verbs are stored with their argument structure specifications and θ-role assignment possibilities. The lexicon is discussed in Dorr (1987), but will not be emphasized in this report.

[8]Hale (1973) describes how this and several other phenomena in Navajo reveal parametric variation of linguistic principles.

2.2 θ-Theory Parameters: Clitic Doubling

θ-Theory is the theory of thematic (or semantic) roles. A principle of this theory is the θ-Criterion which states that each noun phrase argument of a verb is uniquely assigned a semantic role (henceforth θ-role) and each θ-role is uniquely assigned to an argument. For example, the verb *ver* (= see) uniquely assigns a θ-role of *goal* to its direct object:

(11) (i) Juan vio el libro.

(ii) Juan lo vio.

In (11)(i) the *goal* θ-role is assigned to the noun phrase *el libro* (= book) and in (11)(ii) the *goal* θ-role is assigned to the object pronoun *lo* (= him). In order for θ-roles to be assigned to arguments of a verb, there is a principle of θ-role transmission that maps θ-roles in the dictionary entry of the verb to the verbal arguments in the sentence.

In Spanish, the phenomenon of *clitic doubling* is relevant to parametric variation of the θ-role transmission principle. A *clitic* is a pronominal constituent that is associated with a verbal object. For example, the pronoun *le* in the following sentence is a clitic associated with *Juan*, the object of the verb *regalé*:

(12) Le regalé un libro a Juan.
'I gave a book to John.'

In general, a pronominal clitic is associated with a lexical referential NP. Thus, clitic doubling is defined in terms of the pair <*clitic, lexical NP*> where the clitic must agree in number, person, and gender with the lexical referential NP. In (12) the clitic *le* actually stands for an NP that does not yet have a θ-role (namely, *Juan*).

In order to satisfy the θ-Criterion, a parameter of variation is required for the principle of θ-role transmission. Jaeggli (1981) proposes that clitics supply θ-roles to object NPs that are doubled through a θ-role transmission rule:

(13) $[\text{CL } +case_i\ +\theta_j] \ldots [\text{NP } +case_i] \Rightarrow [\text{CL } +case_i\ +\theta_j] \ldots [\text{NP } +case_i\ +\theta_j]$

This rule allows a doubled NP object to receive θ-role as long as the clitic and NP have the same case.[9] If a clitic is not present, a θ-role is assigned in the usual fashion, from the verb that contains the argument in its dictionary entry. Thus, for languages that allow clitics, clitic doubling must be available as a parameter of variation to the θ-role transmission principle of θ-Theory. The θ-Criterion can then be used as a well-formedness condition during parsing so that clitic doubling constructions will be ruled out unless (13) is allowed to fire. This is important in a parsing model since languages that allow clitics could not be analyzed uniformly without such a parameter of variation.

2.3 Trace Theory Parameters: Choice of Traces and Pro-Drop

Trace theory is another subtheory of GB that is important for uniform parsing across languages, in part because it explains the distinctions between languages that allow null subjects (like Spanish) and other languages. A trace is an empty position that is either

[9]Case Theory is not described here. See Chomsky (1981).

base-generated or left behind when a constituent has moved. In this discussion we will talk only about NP traces. However, there may be other types of traces. Thus, the choice of traces for each language is specified as a parameter setting to the trace module.

According to the analysis of the null subject (or pro-drop) parameter introduced by van Riemsdijk and Williams (1986), the choice of whether sentences are required to have a subject is allowed to vary from language to language. In Spanish, as in Italian, Greek, and Hebrew, morphology is rich enough to make the subject pronouns redundant and recoverable. Thus, we can have this sentence:

(14) Hablé con ella.
'(I) spoke with her.'

Since the inflection on the verb is first person singular, the subject pronoun *yo* (=I) need not be used.

The formulation of the *pro-drop parameter* by van Riemsdijk and Williams is motivated by the observation that subjects are missing in a variety of constructions, not just in cases like (14). These constructions do not appear in many other languages (*e.g.*, English, *etc.*); thus, there must be some common factor that will account for the distinction between pro-drop and non-pro-drop languages. The *pro-drop parameter*, then, is a minimal binary difference that does or does not allow empty noun phrases to occupy subject position. (For details on the pro-drop parameter, see van Riemsdijk and Williams, pp. 298–303.) The parameter-setting approach is more desirable than a rule-based approach since it accounts for several types of null subject constructions without requiring several independent rules.[10] The pro-drop parameter is important in the parsing model because it allows uniform analysis of null subject and overt subject languages, ensuring that sentences without a subject are ruled out unless the pro-drop parameter is set.

2.4 Principles and Parameters

Figure 3 contains a table summarizing the subsystems of principles and parameters (grouped according to subtheory) relevant to the parsing model presented here. Because of space limitations, only those parameters that are relevant to a condensed description of the parser are shown. The actual implementation currently has 20 parameters. Figure 4 summarizes the parameter settings required for parsing Spanish and English.

3 Parsing Implementation

The parser is one of three translation stages in an interlingual translation system, UNITRAN (Dorr, 1987), which is implemented in Common Lisp and currently translates simple

[10] A rule-based approach (*e.g.*, Gazdar *et. al.*, 1985) would require a separate rule for every possible null subject construction allowed in a pro-drop language including free subject inversion, relative clauses, that-trace constructions, resumptive pronouns, *etc.* (See van Riemsdijk and Williams (1986) for a discussion of these constructions.) Although GPSG provides a metarule formalism for handling more "top-level" phenomena (*e.g.*, passivization), no generalization is made for closely related phenomena. Furthermore, metarules force the grammar to grow rapidly, thus inducing additional slowdowns during parsing. The parameter-setting approach obviates the need for independent treatment of closely related phenomena without causing a grammar blow-up.

Theory	*Principles*	*Parameters*
$\overline{X}$	A phrasal projection (X^{max}) has a head (X), a specifier, and a complement	Constituent Order, Choice of Specifiers
θ	[CL $+case_i$ $+\theta_j$] ... [NP $+case_i$] $\Rightarrow$ [CL $+case_i$ $+\theta_j$] ... [NP $+case_i$ $+\theta_j$] if language allows clitic doubling	Clitic Doubling
Trace	Null subjects are allowed for pro-drop languages	Pro-Drop
	An empty position may occur where traces are allowed	Choice of Traces

Figure 3: The principles of GB are modularized according to subtheory. Each principle may have one or more parameters associated with it.

Theory	*Parameters*	*Parameter Values*	
		Spanish	English
$\overline{X}$	Constituent Order	spec-head-comp	spec-head-comp
	Choice of Specifiers	V: have-aux; N: det, *etc.*	V: have-aux, do-aux; N: det, *etc.*
θ	Clitic Doubling	applicable and allowed	not applicable
Trace	Pro-Drop	yes	no
	Choice of Traces	N^{max}, Wh-phrase, V, P^{max}	N^{max}, Wh-phrase, V, P^{max}

Figure 4: The parameter settings associated with the principles of GB are allowed to vary from language to language. Here are some of the parameter settings for Spanish and English.

sentences bidirectionally between Spanish and English. In contrast to the transfer approach (*e.g.*, METAL, Slocum, 1984, 1985), the parser and other translation modules are uniform across all languages with respect to their theoretical and engineering basis. (See figure 5.) The transfer approach, on the other hand, requires several parsers and a third translation stage (the transfer stage) in which one language-specific representation is mapped into another. (See figure 6.) Thus, a separate parser must be supplied for each language in the transfer approach, while in the interlingual approach a single parser is used for all languages. The interlingual approach more closely approximates a true universal approach since the principles that apply across all languages are entirely separate from language-specific characteristics expressed by modifiable parameter settings.[11]

In the METAL system, a context-free phrase-structure rule for building a noun stem and an inflectional ending into a noun is shown in figure 7. Although this rule is equivalent to the simple context-free rule `NN` $\Rightarrow$ `NST N-FLEX`, it contains several complex parts: a constituent test that checks the sons to ensure their utility in the current rule; an agreement `TEST` to enforce syntactic correspondence among constituents; a phrase `CONSTR`uctor which formulates

[11]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.

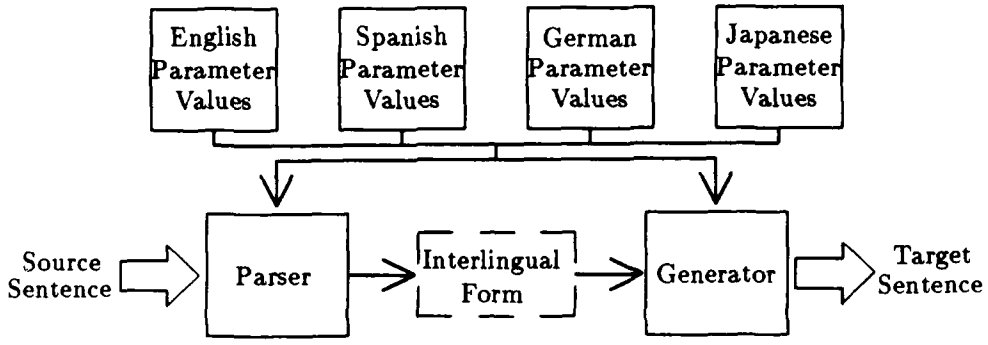

Figure 5: The interlingual design UNITRAN (Dorr, 1987) allows the parser and generator to operate uniformly across all languages.

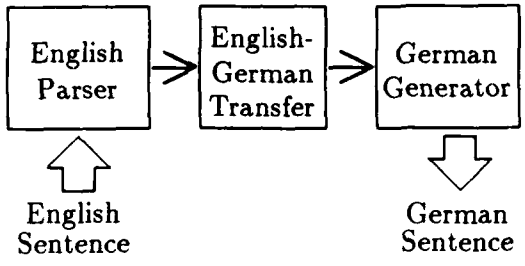

Figure 6: The transfer design of METAL (Slocum, 1984, 1985) requires a separate parser for each language and a transfer component for each source-language target-language pair.

the interpretation defined by the current rule; and one or more target-specific transfer rules. The METAL parser is currently equipped with thousands of such rules.

The UNITRAN parser makes use of parameterized principles rather than hand-written context-free rules to analyze a sentence. The number of parameters is by far smaller than the set of rules that would be needed to handle the same phenomena. The parameters of figure 4 are represented declaratively, and are subject to modification by the grammar-writer. (See figure 8.) There are two types of procedures corresponding to the two boxes of figure 1. The first type includes those procedures that perform structure-building actions (predicting, attaching, and scanning), relying primarily on phrase-structure templates generated at precompilation time. The algorithm that is used to perform these basic parsing actions is the Earley algorithm (see Earley (1970)). The second type consists of constraint verification routines (θ-Criterion, empty NP conditions, *etc.*), performing well-formedness tests on phrase-structures built by structure building procedures.

Before parsing begins, the precompilation stage generates and stores a constant number of underspecified phrase-structure templates per language according to the two $\overline{X}$ parameters of figure 8: constituent order and choice of specifier. When the parser is activated, the structure-building module draws upon these templates, processing each word of input until no more

```
NN              NST             N-FLEX
0               1               2
(LVL 0)         (REQ WI)        (REQ WF)

TEST      (INT 1 CL 2 CL)

CONSTR    (CPX 1 ALO CL)
          (CPY 2 NU CA)
          (CPY 1 WI)

ENGLISH   (XFR 1)
          (ADF 1 ON)
          (CPY 1 MC DR)
```

Figure 7: A context-free phrase-structure rule in the METAL system has a constituent test, a phrase constructor, and one or more target-specific transfer rules. This rule is equivalent to the simple context-free rule **NN** $\Rightarrow$ **NST N-FLEX**, which builds a noun out of a stem and an inflectional ending.

structure-building actions apply. At this time, constraint verification takes place, and the last three parameters of figure 8 are accessed in order to modify or eliminate the structures derived thus far. The parse proceeds in this fashion until all sentence constituents have been successfully scanned, and all constraints have been verified. A sentence is rejected if (a) there is a constraint violation, or (b) after consulting the constraint module, no structure-building actions apply to the remaining input words; otherwise, it is accepted.

Because the constraint module is available during parsing, the phrase-structure templates accessed by the structure-building module need not be very elaborate. For example, a transfer system uses context-free rules of the following form:

(15) S $\Rightarrow$ NP VP
NP $\Rightarrow$ det N
VP $\Rightarrow$ V NP
PP $\Rightarrow$ N PP

However, in the interlingual approach, the very general $\overline{X}$ rule (9) (repeated here as (16) for convenience) subsumes all four of these rules:

(16) $X^{max} \Rightarrow$ (Specifier) X (Complement)

Consequently the grammar size need not, and should not, be as large as those found in other parsing systems. In fact, the number of phrase-structure templates that are generated per language generally does not exceed 150 since there is a limited number of configurations per language that are allowed by the $\overline{X}$ principles accessed at precompilation time. Thus, the

```
(DEF-PARAM CONSTITUENT-ORDER
   :SPANISH (SPEC HEAD COMP) :ENGLISH (SPEC HEAD COMP))

(DEF-PARAM CHOICE-OF-SPEC
   :SPANISH (V (HAVE-AUX) N (DET) I (N-MAX) C (WH-PHRASE))
   :ENGLISH (V (HAVE-AUX DO-AUX) N (DET N-MAX) I (N-MAX) C (WH-PHRASE)))

(DEF-PARAM CLITIC-DOUBLING :SPANISH (T T) :ENGLISH (NIL NIL))

(DEF-PARAM PRO-DROP :SPANISH T :ENGLISH NIL)

(DEF-PARAM CHOICE-OF-TRACES
   :SPANISH (N-MAX WH-PHRASE V P-MAX) :ENGLISH (N-MAX WH-PHRASE V P-MAX))
```

Figure 8: The parameter settings are represented declaratively. The settings for Spanish and English are shown here.

running time of the parser is not subject to the same slow-downs that are found in other systems since the time it takes to search the grammar is reduced.[12]

To clarify the above description of the parsing algorithm, the next section presents an example of how the parsing modules operate.

4 An Example

Consider the problem of parsing (1), repeated here as (17):

(17) Le quiere a Juan
'(She) loves John'

We will look at how the structure-building module determines phrase-structure for this sentence through expansion of non-terminal symbols, and completion of both terminal and non-terminal symbols. At the same time, we will see how the constraint module drops a null subject, processes clitics, and assigns semantic roles. Figure 9 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

[12]It should be noted that the more difficult task is not the grammar search, but the assignment of syntactic analyses to the input. Here, grammar size becomes a more critical issue. There would be a considerable cost if the system were to assign a large number of syntactic analyses to the input before the linguistic constraints had a chance to weed out the incorrect ones. Fortunately, the linguistic constraints apply long before the syntactic analyses have reached completion. In fact, most of the incorrect analyses are weeded out as soon as the offending phrase has been considered by the parser.

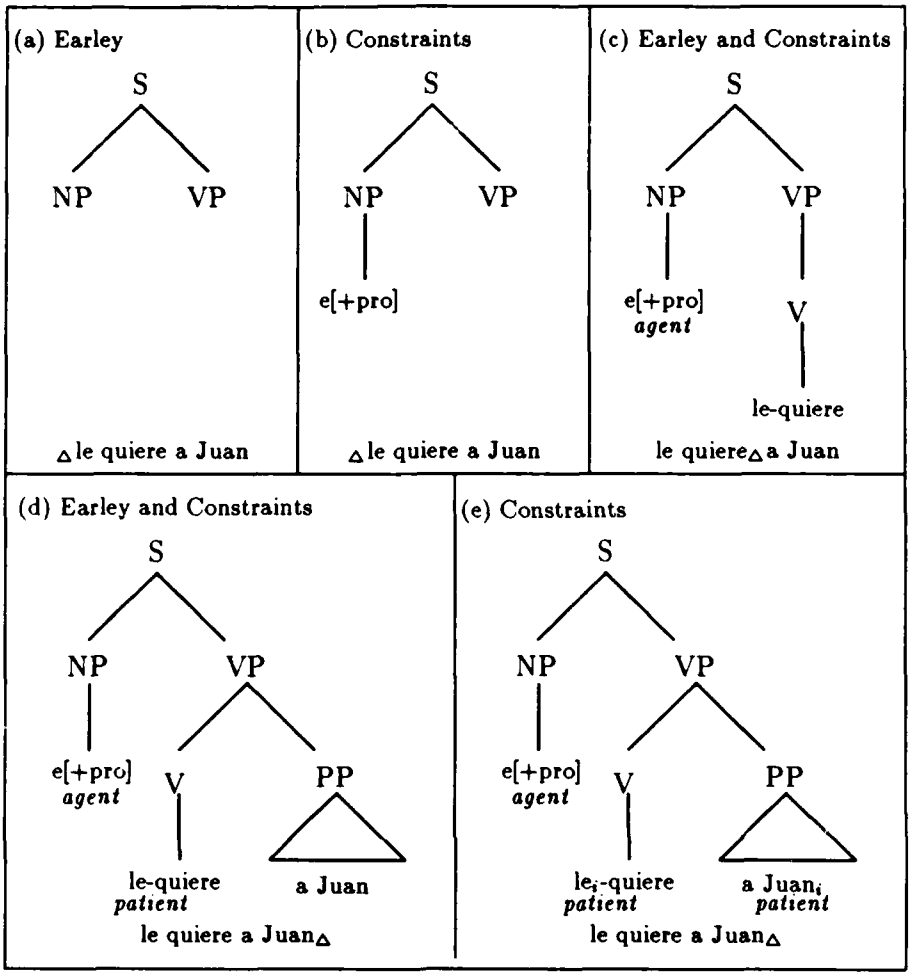

Figure 9: Snapshots of the parser in action show phrase-structure building, null subject dropping, clitic processing, and semantic role assignment for the sentence *le quiere a Juan*.

"constituent order" parameter at precompilation time.[13] 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 (see section 2.3), which dictates that the empty element attached to NP is a subject. The [+pro] (pronominal) feature is associated with the node (see (b)) so the subject will accommodate both null-subject and overt-subject source languages.[14]

[13]Since Spanish is a *head-initial* language, NP must precede VP; however, this would not be the case for non-*head-initial* languages. (See fn. 3 for a description of the "constituent order" parameter.)

[14]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*

In snapshot (c), the Earley module expands VP and scans the first two input words *le quiere*.[15] 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 *quiere* which dictates that this verb requires both an agent (assigned to the subject or *external argument* of the verb) and a patient (assigned to the object or *internal argument* of the verb). The dictionary entry for *querer* (the root form of *quiere*) is encoded as follows: (querer: [ext: agent] [int: patient] V (english: love) (french: aimer) ...)

Now the constraint module predicts that a noun phrase (corresponding to the internal argument of *querer*) must be available. Because the clitic-doubling parameter is set, it is determined that the NP *le* can act as an object of the verb *quiere*; consequently, the NP receives *patient* θ-role as dictated by the lexical entry of *querer*. The constraint module then "records" the fact that a clitic has been seen, so that the NP corresponding to *le* will have a θ-role transmitted to it later if it appears in the input.[16] Once control is passed back to the Earley module, the final two words are scanned, thus completing the PP. Snapshot (d) shows the parse thus far.

At this point the constraint module attempts to assign θ-role to the NP *Juan*. However, all of the θ-roles from the lexical entry of *querer* have already been assigned; thus, assigning a role from this entry would be a violation of the θ-criterion. On the other hand, leaving *Juan* without a role also violates the θ-criterion. Consequently, the constraint module determines (via the clitic-doubling parameter setting) that the θ-role transmission rule (13) is applicable, and recognizes that the NP *Juan* corresponds to the "recorded" clitic preceding the verb *quiere* (since the two match in person, number, and gender). Thus, a θ-role of *patient* is transmitted to *Juan*.[17] As a result of the application of the θ-transmission rule, *le* and *Juan* are coindexed; thus, these two constituents are interpreted as coreferential during the stages following the parse. The final parse is illustrated in snapshot (e).

5 Summary and Limitations

The system described here is based on modular theories of syntax that include systems of principles and parameters rather than complex, language-specific rules. There are three advantages to using the principle-based approach. First, cross-linguistic generalization is captured. The parser operates uniformly across all languages by using general principles that are parameterized according to the language being parsed. The grammar-writer has access to parameters associated with the system principles, thus enabling extension of the system to additional languages.

The second advantage to the principle-based approach is that the grammar size is no

---

or *she* in English) that agrees with the main verb.

[15]Clitic adjunction is generated at precompilation time. The presence or absence of a clitic for a particular language is determined by an adjunction parameter setting associated with $\overline{X}$. This parameter will not be discussed here.

[16]Since clitic doubling is optional, the parse will not be discarded if the corresponding NP does not appear in the input; however, if it does appear (as it does in the above example), it is correctly assigned θ-role.

[17]Note that the θ-role *patient* is assigned to the NP *Juan*, not to the PP *a Juan*; in general, the structural entity that is assigned semantic role is an NP, regardless of the type of phrase containing it.

longer enormous. The presence of linguistic constraints allows phrase-structure templates to be underspecified (more general), thus reducing grammar size for a given language. The reduction in grammar size is crucial for reducing the processing time of the parser.

The third advantage is that the system preserves the modular organization of new theories of grammar. The "co-routine design" of the system divides the tasks of structure-building and linguistic constraint application into two modules. The linguistic constraint module is further broken down into modules associated with each linguistic subtheory. The modularity imposed by the GB framework is an improvement over context-free based systems because it allows general conditions to be factored out, thus simplifying each system component and reducing natural language descriptions.[18]

In summary, the principle-based parsing approach allows uniform parsing across languages, reduces grammar size, and preserves modularity. The approach is an improvement over parsing strategies that limit their coverage and perform poorly due to formidable grammar size. Because of its linguistically motivated basis, the principle-based approach overcomes many of the problems found in rule-based parsing systems.

The primary limitation of the system is that it is almost entirely syntactic-based. The inclusion of θ-roles aids the processing of many semantically equivalent but structurally divergent source and target language predicates. However, the system must be extended to include a more general method of handling structurally distinct but semantically equivalent constituents. Furthermore, disambiguation requiring semantic processing has not been attempted; the extended system should be able to handle semantic disambiguation. A lexical-semantic approach to translation is currently under investigation. The hope is that this new approach will eliminate some of the shortcomings of the entirely syntactic approach.

6 References

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

Bates, M. (1978) "Natural Language Communication with Computers," Springer-Verlag, 191–254.

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

Chomsky, Noam A. (1982) "Some Concepts and Consequences of the Theory of Government and Binding," MIT Press.

[18]There is a subtle difference between the modularity provided by a linguistic formalism and that provided by a programming language. In a modular linguistic system, the surface effects of a change may range far beyond the original source of the change. However, changing a single module does not affect the way other modules *operate*; it only affects the way the modules *interact*. For example, changing principles that determine constituent order does not affect the principles that relate pronouns to their referents, and vice-versa; on the other hand, the direction of pronominal-reference will indeed change from language to language according to how the constituent-order parameter is set. Understanding the distinction between the *operation* of principles and the *interaction* of principles is crucial in order to appreciate the benefit of modularity in a linguistic system. The point is that the different aspects of natural language can be dealt with independently, thus avoiding the task of having to think about the multiplicative surface effects of linguistic principles during the development of the system.

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.

Hale, K. (1973) "A Note on Subject-Object Inversion in Navajo," in *Issues in Linguistics: Papers in Honor of Henry and Renee Kahane*, B. Kachrue *et. al.* (eds.), University of Illinois Press, Urbana.

Jaeggli, Osvaldo Adolfo (1981) *Topics in Romance Syntax*, Foris Publications, Dordrecht, Holland/Cinnaminson, USA.

Robinson, J. J. (1982) "DIAGRAM: A Grammar for Dialogues," *Communications of the ACM* 25:1, 27–47.

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.

Woods, William A. (1970) "Transition Network Grammars for Natural Language Analysis," *Communications of the ACM* 13:10, 591–606.