A Parametric NL Translator

Randall Sharp

Dept. of Computer Science
University of British Columbia
Vancouver, Canada

Abstract

This report outlines a machine translation system whose linguistic component is based on principles of Government and Binding. A "universal grammar" is defined, together with parameters of variation for specific languages. The system, written in Prolog, parses, generates, and translates between English and Spanish (both directions).

1. Introduction

The theory of Government and Binding (GB) (Chomsky 1981) proposes universal principles of grammar (UG) that underlie all natural grammars, with the principles being subject to parametrization, giving rise to language variation. One example is the so-called "pro-drop" parameter of languages like Spanish, Hungarian, Chinese, etc., that, among other things, allow "missing" subjects. To model such a theory, one needs merely (!) to define the UG along with an appropriate set of parameter values for each language, as opposed to defining totally independent language-specific grammars. The overall size of the system is thereby reduced, and adding the grammar for a new language ideally means simply adding its parameter settings. Hence the motivation for a GB approach to machine translation.

To test the approach, a model of UG was constructed, parameters defined for Spanish and English, and language-specific rules added for morphology and, where necessary, idiosyncratic syntax. The result is the GB-Translator (GBT) (Sharp 1985).

A general knowledge of GB is assumed in the following, as is a familiarity with Prolog, the language in which the GBT is implemented.

The overall strategy of the GBT is as follows:

(1)
a. read in a sentence in the source language;
b. perform a morphological and constituent analysis, building an S-structure representation;
c. convert the S-structure to D- ("deep") structure via reverse-transformations;
d. perform lexical transfer from the source D-structure to target D-structure;
e. transform the D-structure to one or more S-structures;
f. validate each S-structure for well-formedness;
g. "flatten" and display only the well-formed target sentences.

Since GB is primarily a theory of syntax, no semantic processing has been introduced in the GBT. What is of interest here is the variety of syntactic forms available in diverse languages for expressing semantically equivalent propositions, with "equivalence" being controlled by appropriate choices of lexical items in the two languages. Problems of semantic ambiguity and structural alteration have not been addressed, as this relates more to translation theory than to GB.

An example of the system's behavior is shown below, where user responses are underlined:

(2)
Enter source language: english.
Enter target language: spanish.
Enter sentence in english: I believe John to have told the truth.
Yo creo que Juan ha dicho la verdad.
Creo que Juan ha dicho la verdad.

Enter source language: spanish.
Enter target language: english.
Enter sentence in spanish: Creo que has dicho la verdad.
I believe that you have told the truth.
I believe you have told the truth.
I believe you to have told the truth.

2. System Structure

The GBT contains a UG component and two language-specific components, one for English, one for Spanish. Exploiting the modular nature of GB (see Wehrli 1983), the UG consists of a phrase structure component based on X-bar syntax (Jackendoff 1977), a transformational component that includes the rule Move Affix (=affix-hopping) and the general rule of Move Alpha (including the subjacency constraint), and a well-formedness component containing constraints on surface representations. The UG, then, is an expression of the theories of X-bar, Case, Theta, Government, Binding, and Bounding (Chomsky 1981) (see Fig. 1).

The language-specific components consist of a lexicon and a grammar. The lexicon includes (1) the lexical entries, and (2) tables of inflections and contractions (e.g. Spanish del = de + el). The grammar contains the UG parameters and idiosyncratic transformations. Figure 2 lists the parameters and transformations currently implemented for each language.

X-bar phrase structure rules
Transformations
Move Affix
Move Alpha
- Subjacency
Well-Formedness Component
- Doubly-filled COMP Filter
- WH-Filter
- Case Filter
- Empty Category Principle
- Binding Conditions

Fig. 1 UG

ENGLISH	SPANISH
Lexicon	Lexicon
Transfomations	Transformations
Subject-AUX Inversion	Verb Preposing
Have-Be Raising	Null Subject
Do Support	
It Insertion	
Complementizer Deletion	
Parameters	Parameters
Pro-drop: no	Pro-drop: yes
Bounding Node: NP,S	Bounding Node: NP,S-bar
Proper Governor: V,P	Proper Governor: V,[+tns]

Fig. 2 Language Components

3. Lexicon

Dictionary entries are represented as Prolog unit clauses:

```
(3)   dict(Lang,Word,Cat,Features).
```

where 'Lang' is the language, 'Word' is the lexical unit, 'Cat' is the syntactic category, and 'Features' contains a list of morphological and syntactic features along with the transfer value. Some sample entries are shown below:

```
(4)  dict(e,put,v,[subcat([n,p]),spanish(poner)]).
     dict(e,believe,v,[subcat(n),subcat(c),sdel(+),
                       spanish(creer)]).
     dict(e,seem,v,[subcat(c),sdel(+),theta(-),
                       spanish(parecer)]).


     dict(s,poner,v,[subcat([n,p]),english(put)]).
     dict(s,creer,v,[subcat(n),subcat(c),
                       english(believe)]).
     dict(s,parecer,v,[subcat(c),sdel(+),theta(-),
                       english(seem)]).
```

The verbs put and poner subcategorize for an NP and PP. Believe subcategorizes for either an NP or a clause, as does its Spanish equivalent, creer. The former has the S-bar Deletion property, while the latter does not, allowing for an infinitival complement for believe but not for creer (see (2) above). Seem and parecer do not assign thematic roles to the subject, indicated by the feature "theta(-)", thus giving rise to sentences such as It seems John has left/Parece que Juan ha salido. Note that since seem and parecer also have the "sdel(+)" feature, they both exhibit subject-raising: John seems to have left/Juan parece haber salido.

Other features in the lexicon include person, number, gender, tense, irregular forms (e.g. go/went), [±wh], [±pronoun], and [±anaphor] (Chomsky 1982).

4. Phrase Structure

The primary phrase structure rules are given in (5), using X-bar syntax and written in the grammar notation of Clocksin and Mellish (1981):

```
(5)   x2(L,C, [[C,F],Spec,X1|Post] ) --->
            specifier(L,C,Spec),
            x1(L,C,X1),
            postadjunct(L,C,Post).

      x1(L,C, [PreHD,HD|PostHD] ) --->
            preadjunct(L,C,PreHD),
            x(L,C,HD,SUBCATS),
            complements(L,SUBCATS,PostHD).
```

For language L and category C, the x2 rule above constructs a parse tree containing nodal information (category and what will become phrasal features) followed by the specifier, the X1 sub-structure, and any post-adjuncts (e.g. PP modifiers). The x1 rule parses a pre-adjunct, the head, and the complements of the head, the latter taken from the subcategorization features of the head. In this way, the parser is head-driven (Proudian and Pollard 1985); the head determines the course of further parsing. The rule in (5) is used for all major phrasal categories, i.e. NP,VP,AP,PP, as well as the clausal phrases COMP (=S-bar) and INFL (=S) (Stowell 1981).

(It should be pointed out that (5) reflects head-initial grammars. A simple parameter could be inserted to accommodate head-final grammars, such as for SOV languages like German, but this introduces parsing problems for Prolog.)

As an example, the following structure is created for the sentence The man had seen many things from his window, where the symbol $e denotes an empty value, and "_" denotes a place-holder for features:

```
(6)   [[c,_], $e, [ $e, $e,
        [[i,_], [[n,_], the, [ $e, man ]], [ $e, $e,
          [[v,_], had, [ $e, seen,
            [[n,_], many, [ $e, things ]]],
            [[p,_], $e, [ $e, from,
              [[n,_], his, [ $e, window ]]]]]]],
        [mode,decl]]]
```

5. System Operation

Following the strategy in (1), the GBT reads in a sentence (assumed to be grammatically correct), analyzes the morphology of each word, and applies the phrase structure rule (5) recursively to build the S-structure. Then, all movement transformations are undone and features percolated to the phrasal node, at which time feature agreement is checked. The result is a D-structure in which all (and only) thematic elements are in θ-marked positions, thereby satisfying the Theta Criterion (Chomsky 1981). This also simplifies lexical translation (as opposed to translating between LF representations).

The transformation stage presents the most interesting aspect of the system, since this is where the principles of UG are applied. (Since the input is assumed to be grammatical, it is not tested for well-formedness.) The high-level Prolog program for this stage is given below:

```
(7)  transformation(L, Dstructure, Sstructure) :-
       transform(L,matrix(+),Dstructure,Sstructure),
       dbl_COMP_filter(L, Sstructure),
       wh_filter(L, Sstructure),
       case_filter(L, Sstructure),
       binding_conditions(L, Sstructure),
       ecp(L, Sstructure).
```

The first action is to transform the D-structure to an S-structure. The "transform" predicate is called recursively on each cyclic node (=S-bar), beginning with the most deeply embedded one. The transformations include those listed in Figure 2 above plus the general transformations of Move Affix and Move Alpha.

The next step involves checking the resulting S-structure for well-formedness. (Note that the well-formedness conditions could execute in parallel, given appropriate machine architecture.) An S-structure that fails to pass any of the conditions forces backtracking into the "transform" predicate. For example, the clause in (8a), which involves no movement, will be generated, but since John cannot be assigned a grammatical Case, it fails the Case Filter. Backtracking to Move Alpha results in John moving to the non-θ-marked subject of seem (8b), resulting in a well-formed structure:

(8) a. *It seems John to have left.
    b. John seems t to have left.

Another example that illustrates how parameters affect generation is given by the Empty Category Principle (ECP), which requires traces to be properly governed. Given the following parameter settings:

```
(9)  proper_governor(english,v).
     proper_governor(english,p).

     proper_governor(spanish,v).
     proper_governor(spanish,i).
```

where the last statement is interpreted as "INFL is a proper governor if it contains the feature [+tns]", the "ecp" statement in (7) will allow preposition-stranding in English but not in Spanish (10), and allow "that-trace" in Spanish but not in English (11):

(10) Which film did they leave after t?
     After which film did they leave t ?
     Después de cuál película salieron ellos t ?
     *Cuál película salieron ellos después de t ?

(11) Who seems t to have left?
     Who does it seem t has left?
     *Who does it seem that t has left?

     Quién parece t haber salido?
     Quién parece que t ha salido?

A similar use of parameters controls subjacency within Move Alpha. To show that S-bar is a bounding node in Spanish, Torrego (1984) notes that Verb Preposing must occur in every clause that contains a wh-phrase or its trace in COMP. In (12a), the trace of con quién causes inversion whereas (b) is derived without movement to COMP, obviating inversion:

(12)
a. Con quién sabía Juan t que había hablado María t ?
b. Con quién sabía Juan que María había hablado t ?
   ('With whom did John know that Mary had spoken?')

6. Conclusion

The GBT system operates solely on the basis of syntax. In a more complete translation system, issues of semantics, pragmatics, and discourse must be dealt with, ideally by again assuming general principles subject to parametric variation. Nevertheless, the current system illustrates the feasibility of a generalized syntactic component in an overall language processing device.

7. Acknowledgements

This paper is dedicated to the memory of Alfredo Hurtado.

References

Chomsky, N. (1981) Lectures on Government and Binding, Dordrecht: Foris Publications.
_____ (1982) Some Concepts and Consequences of the Theory of Government and Binding, MIT Press.
Clocksin and Mellish (1981) Programming in Prolog, Berlin: Springer-Verlag.
Jackendoff, R. (1977) X-Bar Syntax: A Study of Phrase Structure, MIT Press.
Proudian, D. and Pollard, C. (1985) "Parsing Head-Driven Phrase Structure Grammar", 23rd Annual Meeting of the ACL, 167-171.
Sharp, R. (1985) A Model of Grammar Based on Principles of Government and Binding, M.Sc. Thesis, Univ. of British Columbia, Canada.
Slocum, J. (1985) "A Survey of Machine Translation: Its History, Current Status, and Future Prospects", Computational Linguistics, 11:1-17.
Stowell, T. (1981) Origins of Phrase Structure, Ph.D. dissertation, MIT.
Torrego, E. (1984) "On Inversion in Spanish and Some of its Effects", LI 15:103-129.
Wehrli, E. (1983) "A Modular Parser for French", Proc. of 8th IJCAI, 686-689.