Generic Parsing and Hybrid Transfer in Automatic Translation

Christopher Laenzlinger, Sébastien L'haire & Juri Mengon

LATL, University of Geneva

Abstract. The ITS-3 project (LATL, University of Geneva) aims at developing an automatic translation system involving French, Italian, English and German. The translation system relies on the classical architecture parsing-transfer-generation. Parsing of the source language is done by the IPS system, which is based on the Principles & Parameters Theory of Chomsky's Generative Grammar (Chomsky and Lasnik, 1995). The parser produces rich syntactic structures containing lexical, phrasal, grammatical and thematic information, and focuses on (i) robustness, (ii) genericity, and (iii) deep linguistic analyses. These properties are essential for a system to be efficient in multilingual large-scale applications. The transfer mechanism acts on hybrid structures, called *pseudo-semantic structures* (PSS), that mix lexical items with abstract semantic information. On the basis of these PSS, the generation module produces correct output sentences

Introduction

The ITS-3 project (LATL, University of Geneva) aims at developing an automatic translation system involving French, Italian, English and German. The translation system relies on the classical architecture parsing-transfer-generation. Parsing of the source language is done by the IPS system, which is based on the Principles & Parameters Theory of Chomsky's Generative Grammar (Chomsky and Lasnik, 1995). The parser produces rich syntactic structures containing lexical, phrasal, grammatical and thematic information, and focuses on (i) robustness, (ii) genericity, and (iii) deep linguistic analyses. These properties are essential for a system to be efficient in multilingual large-scale applications. The transfer mechanism acts on hybrid structures, called *pseudo-semantic structures* (PSS), that mix lexical items with abstract semantic information. On the basis of these PSS, the generation module produces correct output sentences. We will illustrate how the translation works with German as the source language and French as the target language.

1 Using a Principle-based approach

Rule-based grammars, such as the context-free backbone of GPSG (Generalized Phrase Structure Grammar, Gazdar et al. 1985) or LFG (Lexical Functional Grammar, Bresnan, 1982), mainly use context-free phrase structure rules to describe the surface pattern of a language. Parsers relying on these architectures have undeniable advantages due to their well-known mathematical and computational properties. These lead to a uniform description of the grammar and therefore make it easier to calculate their run-time complexity. Furthermore, there are several efficient parsing algorithms available for grammars expressed with phrase structure rules. Despite these advantages rule-based grammars often also have to face serious shortcomings. The most important limitations are due to the fact that their nature is construction specific and language-dependent. Thus, moving towards a multilingual implementation means that the rules in question need either to be expanded automatically or have to be multiplied by the number of languages treated by the system.

To face the problem of the construction-specific and language-dependent nature of phrase structure rules, other formalisms rather prefer to use constraints to restrict the number of phrase structure rules. The unification and constraint-based approach is realised, i.e. in the HPSG (Head-driven Phrase Structure Grammar, Pollard & Sag, 1994) formalism. A different approach is realised within Principles & Parameters Theory (Chomsky, 1981; Chomsky and Lasnik, 1995). Within this framework, grammar is conceived as a set of interactive principles

of well-formedness, which hold cross-linguistically, and as a set of language-specific parameters (see also Berwick, 1991). As we will see in the following sections, the principle-based framework has proven to be a useful model for the implementation of a large-scale multilingual translation system.

2 IPS : the parsing system

IPS is a principle-based parsing system which differs from rule-based parsers in its modular architecture. The principles of the grammar are implemented as generic modules which dispense with phrase structure rules. These modules hold for all languages. Other modules realize language specific properties, corresponding to the values of the parameters.

The core generative module is the phrasal X' module that rules the general geometry of syntactic structures. All constituents are formed according to the X' format in (1).

(1) $[\text{Specifier}_{\text{list}} \quad X° \quad \text{Complement}_{\text{list}}]_{XP}$

Due to simplicity, the bar-level is not represented, while Specifier and Complement are implemented as (eventually empty) lists of maximal projections. X° stands either for a lexical head (Adv, Adj, N, V, P) or for functional category (C, I, D, F)[1], which all project a maximal projection (XP). The uniformity of phrasal projections is obtained by the category-independency of the X' schema implemented in the IPS parser. Consider the position of verbal complements in German. Objects can either precede or follow the verb depending on their category (nominal/prepositional vs. sentential). Hence, they can occur either in the Complement list on the right of the verbal head or in the Specifier list on the left of this head.[2] Actually, the attachment module specifies which type of constituents can be attached in the Complement or the Specifier list of a specific head. The attachment procedure builds configurations determined by properties of selection from heads and filtered by agreement relations. Further relationships between constituents are construed through the chain-building module. Formally, the parser inserts traces into a syntactic structure and tries to bind them to their potential antecedents. As general devices, the X' schema and the attachment module act as generating operations. To avoid overgeneration and ungrammatical parses, the IPS parser makes use of top-down filtering constraints, such as the thematic module, establishing thematic/semantic relations between a predicate and their arguments (agent, theme, goal, etc.), and the case assignment module, which requires that each lexical nominal phrase be associated with a morphological or abstract case (nominative, accusative, dative, etc.).

Among language-specific modules for German, the IPS parser applies the verb second (V2) constraint, the Object-Verb (OV) vs. Verb-Object (VO) ordering conditions, the constituent reordering rules ('scrambling'), and some other constraints on Germanic specific constructions (extraposition of relatives, *infinitivus pro participio*, coordination possibilities, etc.).

We will illustrate how the parsing mechanism works with the following example.

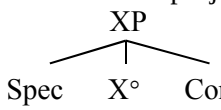
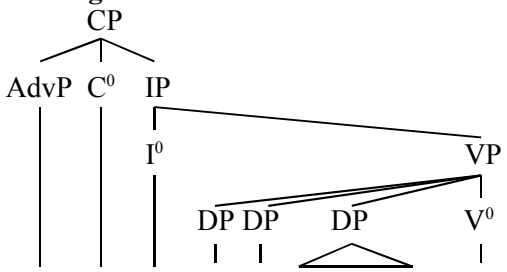
(2) Dann hatte Hans es dieser Frau geschenkt.

[1] The following abbreviations are used to represent the constituents: Adj(ective), Adv(erb), N(oun), V(erb), P(reposition), D(eterminer), C(omplementizer), I(nflection) and F(unctional).

[2] In this sense, the Specifier and Complement lists do not have an interpretative function as such. Thus, a true verb complement can occur in a Specifier list, while an adjunct (i.e. a specifier-like element) can occur in a Complement list.

Then had Hans it this woman offered
'Then John had offered it to this woman.'

The IPS parser first undertakes a lexical analysis by segmenting the input sentence into lexical units and inserts them as edges for all their possible readings into a chart (Table 1, step 1). For every lexical unit in the chart, a maximal projection is provided and lexical features are transferred to the syntactic level (XP) in accordance with the Projection Principle (Table 1, step 2). Then, the attachment module checks all possible attachments to the Spec and Compl positions of the various XPs[3] according to selectional rules and thematic relations. The backbone of the clause structure is CP > IP > VP, as shown in Table 1, step 3. IP functions as the right complement of CP, while VP is the right complement of IP. In main clauses, the first constituent, here the AdvP *dann,* occupies the Spec-CP position. The tensed verb is raised to C°, the second position of the clause. This is a simple way of applying the V2 constraint. Since tensed auxiliaries are base-generated in I°, the chain formation module links the raised auxiliary in C° with its trace in I° (Table 1, step 4). In our analysis, the DP subject *Hans*, is attached in Spec-VP. The OV configuration of German results from the attachment of the DP[4] complements *es* and *seiner Frau* in Spec-VP. The past participle *geschenkt* occupies the V° position. At this point, the filtering interpretative modules verify that each argument of the clause be assigned a thematic role from the verb (thematic module), while every nominal phrase be in the right configuration to receive a case (case module, Table 1, step 5). The whole syntactic structure of the input sentence is given in Table 1, step 3.

Steps	Process
1	**Lexical analysis:** [$dann_{Adv}$] [$hatte_{I}$] [$Hans_{D}$] [es_{D}] [$dieser_{D}$] [$Frau_{N}$] [$geschenkt_{V}$]
2	**Projection:** Project a maximal projection XP for every element in the lexical analysis chart  where X° ∈ {N, V, A, P, D, Adv, C, I}
3	**Checking of attachments:** 
4	**Chain formation:** (hatte, t) $_i$
5	**Argument interpretation:** Arg 1: *Hans* CASE = nominative; θ-ROLE = agent Arg 2: *es* CASE = accusative; θ-ROLE = theme Arg 3: *dieser Frau* CASE = dative; θ-ROLE = beneficiary

[3] The IPS parser is a bottom up licensing parser that pursues all alternatives in parallel. Its parsing strategy is called "right-corner attachment" (Wehrli, 1997: 220). A more detailed description of the parsing algorithm used in IPS can be found in Wehrli, 1997).

[4] We follow Abney's (1987) DP-hypothesis according to which D is the head of nominal expressions and takes an NP as its complement.

Table 1: Parsing process

In addition to modularity and genericity, the IPS system is characterized by two other important properties: robustness and the use of rich and deep linguistic analyses. Robustness is required for any system to be efficient in large-scale domain-independent applications. For this aim, the IPS parser can treat unknown words and includes micro-grammars (idioms, parentheticals, temporal expressions, etc.). Robustness is increased by the "no-failure" parsing strategy according to which incomplete analyses are still exploitable through partial parsing results. In addition, the use of rich and deep linguistic information allows the parser to be involved in a large number of multilingual domain-independent applications, notably in automatic translation. This approach to language processing contrasts with the quite wide-spread approach adopted by shallow parsers or NP-identifiers which consists of using some kind of linguistic pre-processing adapted to specific applications.

3 The ITS-3 Translation Project

The ITS-3 project (Interactive Translation System, Etchegoyhen and Wehrli, 1998) aims at developing a translation tool using abstract interface structures called *pseudo-semantic structures* (PSS). The PSS present a hybrid nature combining abstract semantic representations with lexical items, and constitute the entries to the syntactic generator GBGen (Etchegoyhen and Wehrle, 1998). To understand how the entire translation procedure works, we will pursue the translation into French of the German example (2). The first step following the syntactic analysis consists in transferring the parse results (Table 1, step 3) to the interface PSS. A PSS contains information about the clause, namely its mood ('real' or 'unreal'), its tense and aspect specifications. The continuum relationship between Reichenbach's (1947) Event time (E) and Speech time (S) values[5], intermingled with aspectual specifications (progressive, perfective, imperfective), are used to determine the tense information in the PSS. Further information about the voice, negation and the utterance type of the clause is specified. Since the PSS involve lexical transfer, which is restricted to open class lexical categories such as verb, noun, adjectives and adverbs,[6] the predicate is specified as a lexical entry. Every PSS can have a certain number of 'satellites' that depend on it. Thus, non-clausal arguments are represented as DP-structures (DPS) of operator-property type, corresponding to the syntactic determiner-noun relation. AdvPs are represented in so-called characteristic structures (CHS). Thus, for sentence (2), a PSS like (3) is derived.

(3) Pseudo-Semantic Structure:
Information about the clause
Mood : real (= indicative)
Tense : E<S (= past)
Aspect : (non progressive, perfective)
Voice : active
Negation : not negated
Utterance type : declaration
Predicate : *schenken* ⇔ *offrir*

Information about the satellites

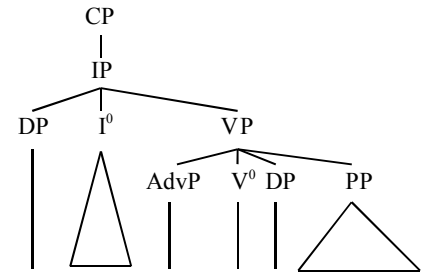
5 The possible values are E<S, E=S, E>S.

6 The use of lexical transfer seems at present unavoidable in automatic translation, provided that the assignment of abstract, lexically-independent, values to open lexical categories is too complex, often inconceivable, to be computed efficiently (see also Vauquois and Boitet, 1988).

1.) DPS	2.) DPS	3.) DPS	4.) CHS
Theta role: agent	Theta role: theme	Theta role: beneficiary	Value: when
Property: *Hans*	Property: Δ	Property: *Frau* ⇔ *femme*	Scope: sentential
Operator: Δ[7]	Operator: Δ	Operator: demonstrative	Characteristic: *dann* ⇔ *ensuite*
	Gender: neutral	Number: singular	
	Person: 3rd person		
	Number: singular		

The semantic representation given in (3) constitutes the entry to the GBGen generator. According to this information, the output sentence will be a declarative, active, non negated clause. The tense will correspond to the French 'indicatif plus-que-parfait' (indicative past perfect). The verbal predicate *offrir* takes three arguments and a sentential temporal modifier. Since the external argument ('agent') is generated as the subject of the clause, it will be realized as a DP attached in Spec-IP. The second argument is a direct object personal pronoun, which will have to be cliticized to the auxiliary in I° and will be linked to a trace in its base position, Compl-VP. The third argument will be realized as a 'dative' indirect object with the subcategorized preposition *à,* and will be expressed as the PP *à cette femme.* This prepositional complement will be attached to Compl-VP. Finally, the fourth satellite will be syntactically generated as a AdvP attached to Spec-VP. The resulting sentence and structure are given in (4).

(4)



Hans l_i*'avait ensuite offert* t_i *à cette femme.*
'Then John had offered it to this woman.'

Conclusion

We have described an automatic translation system based on the classical architecture parsing-transfer-generation. We have illustrated the way in which the system works with German as the source language and French as the target language. Parsing is undertaken by the generic IPS system which provides detailed linguistic analyses. The transfer component uses hybrid lexico-semantic information structures, called *pseudo-semantic structures* (PSS), combining lexical transfer with abstract functional and semantic information. This mixed transfer technique takes advantage of both the simplicity of the lexical transfer procedure and the abstractness of the *interlingua* approach. The generation module takes the PSS as input and gives back correct output sentences.

References

Abney, S. (1987) *The English Noun Phrase in its Sentential Aspect.* Ph. D. thesis. Cambridge, Mass.: MIT Press.

[7] The symbol Δ indicates that the values assigned here are left un(der)specified.

Berwick, R. (1991) "Principles of Principle-based Parsing". In R. Berwick, S. Abney, C. Tenny ed., *Principle-Based Parsing: Computation and Psycholinguistics*. Dordrecht: Kluwer Academic Press.

Bresnan, J. ed. (1982) *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press.

Chomsky, N. (1981) *Lectures on Government and Binding*. Dordrecht: Foris Publications.

Chomsky, N. and Lasnik, H. (1995) "The Theory of Principles and Parameters". In N. Chomsky, *The Minimalist Program*. Cambridge, Mass.: MIT Press, 13-127.

Etchegoyhen, T. and Wehrle, T. (1998) "Overview of GBGen : A Large-Scale Domain Independent Syntactic Generator". In *Proceedings of the 9th International Workshop on Natural Language Generation*. Niagara Falls, 288-291.

Etchegoyhen, T. and Wehrli, E. (1998) "Traduction automatique et structures d'interface". In *Actes de TALN'98*, Paris, 2-11.

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

Reichenbach, H. (1947) *Elements of Symbolic Logic*. New York: Free Press.

Vauquois, B. and Boitet, Ch. (1988) "Automated Translation at Grenoble University". In J. Slocum, *Machine Translation Systems*. Cambridge: Cambridge University Press.

Wehrli, E. (1997) L'analyse syntaxique des langues naturelles. Problèmes et methodes. Paris: Masson.