# EVOLUTION OF THE LOGOS GRAMMAR: SYSTEM DESIGN AND DEVELOPMENT METHODOLOGY

Patricia Schmid and Claudia Gdaniec
Logos Corporation
200 Valley Rd., Suite 400
Mt. Arlington, NJ 07856
{pschmid,cgdaniec}@logos-usa.com

## Abstract

The history of MT shows how the shortcomings of one approach become the catalyst for a new approach. New systems are designed in reaction to old systems' weaknesses. Yet some old systems have survived and are being used increasingly in the "real world" of translation. One of these systems is Logos. While keeping its fundamental approach, Logos has evolved through its own design history so that today's grammar looks quite different from the earliest model. Our paper describes how the Logos system evolved in response to requirements of development, both linguistic and methodological. We describe four "generations" of the Logos English-source system, each situated in its technological context, with the most recent generation allowing for completely separate development of analysis and transfer components. The system has become larger and more complex over the years, but at the same time more modular. Combined with the advances in computer technology, this modularity makes it easier to develop and improve grammars. Linguists can now work independently, both on-site and remotely, on different source and target languages.

## 1. Introduction

The history of MT shows how the shortcomings of one approach become the catalyst for a new approach. The MT field expands as new systems are designed in reaction to old systems' weaknesses. Yet even as the field continues to grow and change, there are some old systems which have not only survived, but are being used increasingly and successfully in the "real world" of translation. One of these commercial systems is Logos. Over the years Logos has evolved through its own design history so that today's grammar looks quite different from the earliest model.

Logos has never fit exactly into one of the traditional MT classification schemes. At one time, a Logos system was designed for a single language pair, on which grounds Warwick (1987:24) groups Logos among the direct systems, but the present system uses a single source analysis as the basis for the generation of multiple targets.[1] Furthermore, unlike a direct system, Logos has always maintained a separation of linguistic data and processing algorithms. Logos also creates a full parse of the sentence, which is characteristic of a transfer system. The SAL representation

[1] Currently Logos has an English-source system with German, French, Spanish, and Italian targets, as well as a German-source system with English, French, and Italian targets.

into which source text is translated is a type of interlingua, which further complicates classification. In addition, the Logos rulebases have been compared to a neural net because of the way they interact with the linguistic input. (Scott 1992).

Thus it cannot be said that Logos has evolved from one MT category into another. It continues to be a hybrid system. Nevertheless, the requirements of linguistic development have over the years led to new stages in the evolution of Logos, both in system design and in development methodology. These requirements have included:

- "unsolvable" linguistic problems
- creation of new language pairs
- the need to improve the speed and efficiency of development.

The developer's goal is to make the MT system more sophisticated linguistically. At the same time, the developer wants to keep the system from becoming overly complex, in order to keep the cost of maintenance and improvement from becoming prohibitive. But improving a system's linguistic capabilities usually means increasing its size and complexity - which adds to the developer's methodological burden. In response to the methodological problems, new ways of working are introduced.

The evolution of Logos has occurred within the context of ever-improving technology, both hardware and software. In this paper we will look at the history of linguistic development at Logos, focusing on the interaction of linguistic improvements, system design, and development methodology. We will describe four "generations" of the Logos English-source system, each situated in its technological context.[2]

Only the first generation represents a completely new Logos system. Each subsequent generation emerged from its predecessor in response to demands for improving translation and improving development. All versions of Logos have been based on the Logos-internal Semanto-Syntactic Abstraction Language (SAL). An input text is translated into an SAL string, and grammar rules at all levels of generality are written in SAL. This original design is the foundation for Logos's homogeneous, self-organizing rulebases and self-determining order of rule application. It also allows the rulebases to increase in size without affecting system performance; thus linguistic coverage can be continually improved through the addition of rules which address new phenomena.

## 2. First Generation (1970-1973)

Figure 1 shows the architecture of the earliest Logos English-source system. This was the English-Vietnamese system, begun in 1970. Between Dictionary Lookup and Generation there were three processing modules. RES was responsible for disambiguating homographs. The TRANslation modules created a rough parse of the English sentence, with TRAN1 looking for low-level syntactic components and TRAN2 working at a higher level. As constituents and partial constituents were identified in the source language, their translations were added to a target-language structure. Each matched TRAN rule performed source and/or target operations as

[2] Space limitations prevent us from describing advances in the design of the Logos dictionary or advances in acquisition tools.

appropriate. Thus, source-language analysis and target operations were interwoven. Both source and target structures were passed from one TRAN to the next, and the final target information was used by the Generation module.

This system was built on a mainframe. The rulebases were maintained on punch cards. Each time a change was made, the entire rulebase had to be recompiled. This system presented developers with several problems. Linguistically: semantics was not handled well; homograph resolution was weak; the parsing of complex sentences needed improvement. Methodologically: the development process was slow due to the technology available at the time.

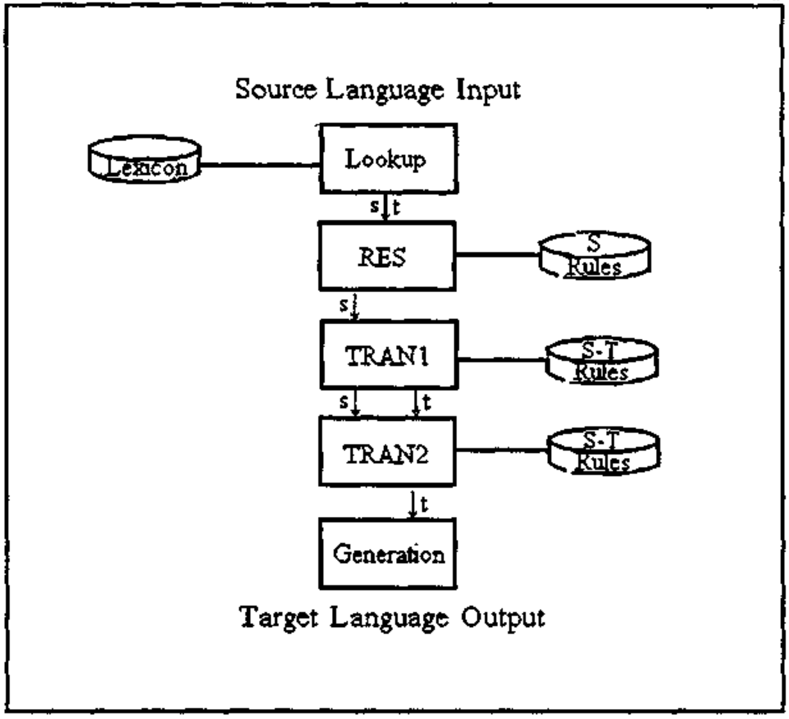

**Figure 1**

## 3. Second Generation (1974-1980)

Figure 2 shows the architecture of Logos during its second phase. By this time, SAL had been refined and expanded, which allowed the creation of more refined rules with slightly better handling of semantics. During this time, three separate systems were built: first English-Russian, then English-French, then English-Farsi. These systems used the same Lookup and RES modules, but each had its own set of TRANs, adapted from the TRANs of the previously built system. Now there were four TRAN modules, which allowed for a more accurate, bottom-up parse of the sentence, including better handling of complex structures. The TRANs had to be separate for each language pair because source (analysis) and target (transfer) actions were still combined in a single TRAN rule.

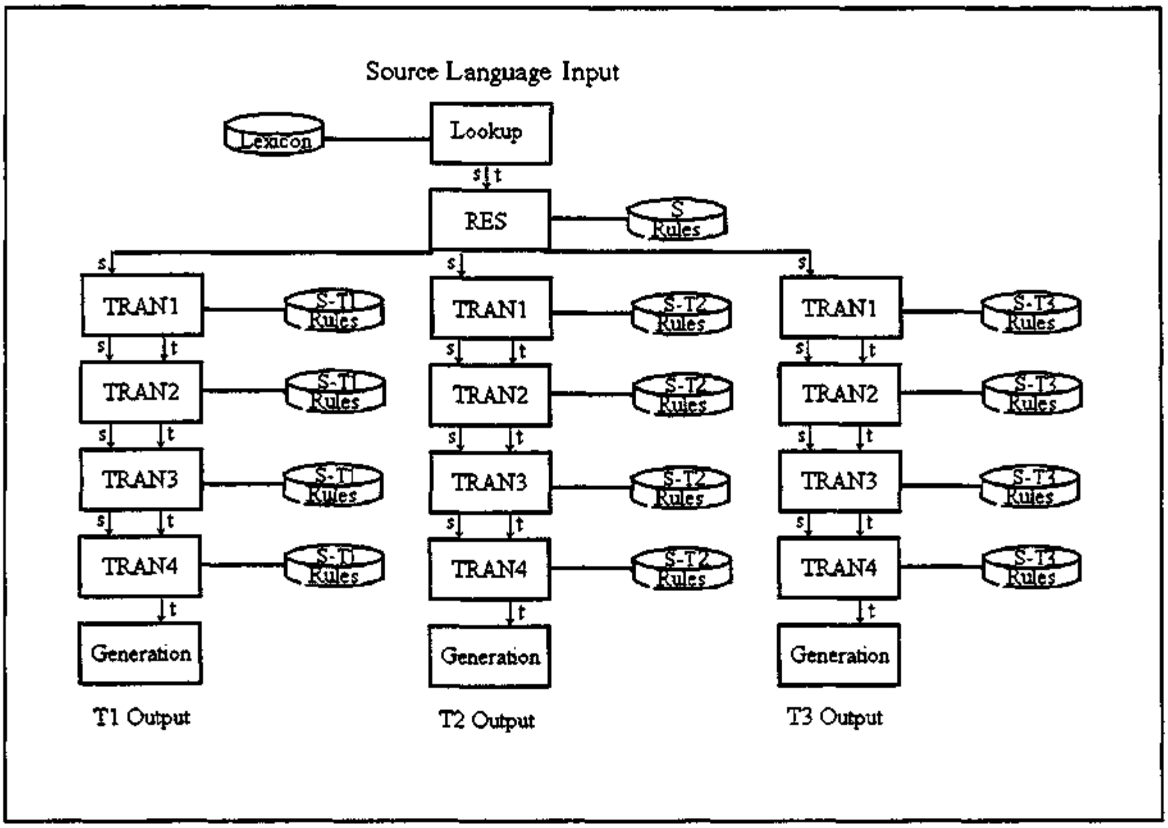

**Figure 2**

At this stage developers had advanced to using dumb terminals and a very inconvenient editor instead of punch cards. But they were still working directly on the rulebase, which took time to compile and recompile. As the system gained modules, it became possible, even necessary, for more linguists to work on different parts of the system. But it was difficult for several people to change the same rulebase without disrupting each other's work. And since the rulebases interacted with each other, one developer's work on one rulebase might conflict with another's work on a different rulebase. By the time the results were seen in the output, all rulebases had been recompiled, and the tedious process of pulling rules back out of the rulebase had to begin. It became obvious that it was not optimal to work directly on the rulebase.

Though linguistically more sophisticated, this version of Logos still did not solve all the problems it faced. Homograph resolution, treatment of semantics, and parsing of complex sentences were all improved, but overall, source analysis was still fairly weak. It also became apparent that maintaining different sets of TRANs for different language pairs meant the duplication of source analysis effort. It would be more desirable to have a multi-target system.

## 4. Third Generation: (1981-1996)

The third-generation phase has spanned a period of fifteen years. During the early years of this phase, the Logos system took several major steps forward both in system capabilities and in development techniques. The resulting system has provided the framework for linguistic development at Logos up to the present day. Figure 3 illustrates the third-generation system architecture. This system contains three innovations which were incorporated in response to challenges faced by the earlier models.

First, a second RES module was added. This module creates a "macro parse" showing the clause structure of the sentence and uses this top-down information to complete the homograph resolution process begun in RES1. This new rulebase gives the system much more accurate homograph resolution capabilities. It also provides the TRAN modules with a global picture of a sentence's structure.

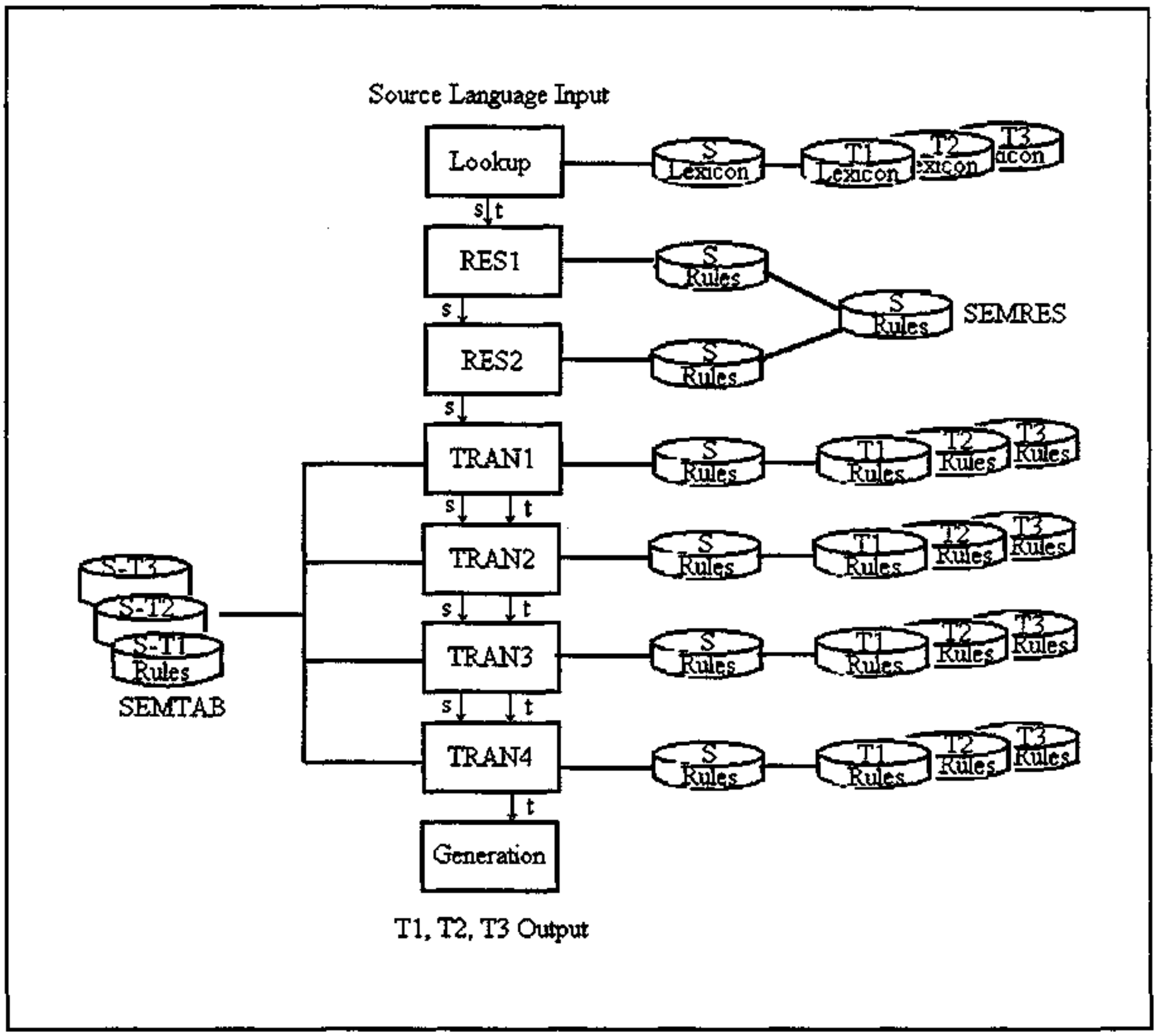

**Figure 3**

Second, in response to the system's deficiencies in handling semantics, semantic rule tables were introduced. RES has the SEMRES table, containing selectional constraint rules which help in disambiguation. The TRANs make use of SEMTAB, which differentiates specific senses of words, contains collocational information,' and specifies translations overriding those from the dictionary. The TRANs consult SEMTAB when making parsing decisions as well as for target-specific information. The addition of SEMTAB has resulted in significant improvements in translation.

Finally, during this stage Logos has become a true multi-target system. During the 1980s, the TRAN rules were revised so that they were no longer a mixture of source and target actions. Each TRAN module now accesses a set of source-only rules. Where the rule has implications for the target, there is a call to a target rule. Each language has its own set of target rules for each TRAN.[3] The source rules operate on the source-language data structures, while the target rules operate on the target structures. Both sets of information are passed from one TRAN to the next. As before, the target information is the input to the Generation module. The separation of source and target rules opened the way for development of new targets for existing source language systems. The English-Spanish and English-Italian systems (as well as German-French and German-Italian) have been developed entirely within the new architecture.

During this phase, the technology with which the system is developed has progressed from mainframe to networked UNIX workstations. Each workstation offers the potential for an individualized work environment. Each linguist has an individual “test” copy of the dictionary and SEMTAB. (All tested, acceptable changes are applied to the central dictionary and SEMTAB databases.) At the same time the linguistic development environment has been redesigned so that each linguist has a private front-end to the rulebases. The design of this environment, as it is implemented for the TRANs, is pictured in Figure 4. Linguists no longer work directly on the rulebases. Instead, each linguist has his/her own individual set of rulebases, called “minis”, each corresponding to a TRAN module. Source and target rules are added, changed, and deleted in the minis. During translation, the system consults first the mini, then the main rulebase for rule matches. While working, the linguist needs only to compile a mini instead of an entire rulebase. This decreases the time required for development. It also allows linguists to develop strategies without immediately interfering with each other's work. Along with the mini configuration, the development environment includes a "merge" procedure whereby all the minis are combined and tested as a whole, then merged into the main system. Individual work environments, in conjunction with the advent of the Internet, have also opened up the possibility of remote development. The French and Italian targets are currently being developed remotely.

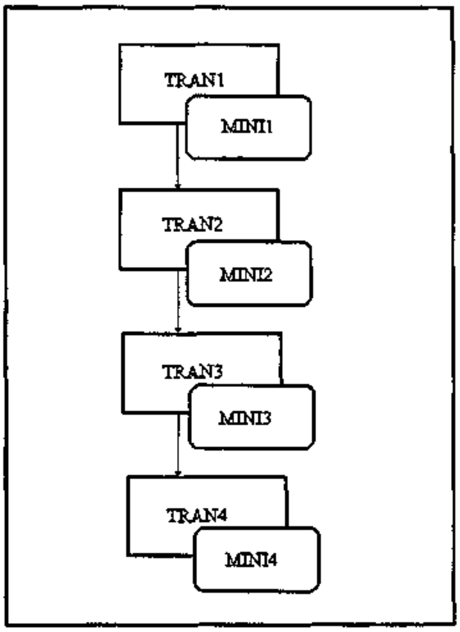

**Figure 4**

A single-source, multi-target system developed in an environment in which linguists work separately from each other provides a rational way for source- and target-language specialists to

[3] In fact, the target rules are separated into three types of "tables." The first type of table contains the target component required by the source rule. The second type contains reusable code. The third type of table contains source-independent target information, and as such has the potential for use with any source language. The Italian target was the first one created according to this design. It thus includes a true Italian-target module which is used in both the English-Italian and the German-Italian systems.

work on separate components of the system. Nevertheless, at least two problems remain, one of them linguistic, the other methodological.

The linguistic problem: Because the target rules are still directly attached to the source analysis, linguists have found that it is sometimes necessary to make decisions about how a target structure should be generated before there is sufficient information available about the source. For example, when a noun phrase is identified in TRAN1, the accompanying target rules are required to specify how the phrase is to be translated. But the phrase *any books* has different possible translations in German, depending on other elements in the surrounding clause. The phrase *May 1st* in the sentence *They chose May 1st* has a different translation from that of *May 1st* in *They voted May 1st*. In these and other cases the target rules need more information about the source sentence in order to make correct transfer decisions.

The methodological problem: The linguists can develop TRAN rules independently of each other, but these rules must still be reconciled with each other. Due to the connection between source and target rules, any change to source rules affects all targets. In general, the source work is beneficial to all targets. But problems can arise during the mini reconciliation phase. The linguist for one language pair may discover that the "source analysis" done in another linguist's mini is really not the kind of analysis that is useful for his/her target. For example, in the analysis done in the English-French mini, the relative clause and its preceding head noun are combined into a single noun phrase. But this parse is not adequate for English-German, because this target still needs to see the relative clause apart from its head in TRAN4 in order to be able to order a sentence's clauses for the German output. A conflict of a different kind occurs in the parsing of postnominal adjective phrases, e.g., *the coins found in the river.* The English parse adds a dummy relative pronoun and *be* in order to make the adjective phrase look like a relative clause. This is very helpful for the German target, which may produce the translation *die Münzen, die im Fluß gefunden wurden,* but causes unnecessary work for French, which produces the translation *les pièces trouvées dans la rivière.*

Of course this sort of problem leads one to the theoretical distinction between "pure" source analysis and analysis from the perspective of transfer. It is easy to bias the source parse in favor of one target or another; within the system shown in Figure 3, one might ask if it is possible to have a pure source parse.

Linguists working within the third-generation model have found that they are spending too much time trying to keep in step with other targets. This experience shows that the development methodology can still be improved. The solution to both the linguistic problem and the methodological problem turns out to be the same one: a rigorous separation of source and target.

## 5. Fourth Generation (under development)

The fourth-generation Logos system, shown in Figure 5, is currently being developed.[4] The four TRANslation modules are being replaced by four PARSE and four TRANsfer modules.[5] We are creating the PARSE rulebases by copying the third-generation TRAN rulebases, then deleting all

[4] It is expected to become part of the general development environment in 1997.

[5] Note that the name *TRAN* has taken on a new meaning.

the linked target rules. The remaining rules are being revised so that their sole function is to create a representation of the English parse. The parse information is passed to TRAN1. There is a separate set of TRAN rules for each language pair, also based initially on the TRAN rules of the previous model. Complete source information is available to them from the start, so the TRAN modules can focus on transfer. The TRAN rules are gradually being revised so that they make use of information sent from PARSE. Where source information is too detailed for one target's needs, the TRAN linguist can ignore it; where it is too general, the TRAN linguist can write the rules needed for extracting more specific information. It is expected that the TRAN rules will evolve into language-pair specific transfer modules which can be developed independently of each other.[6]

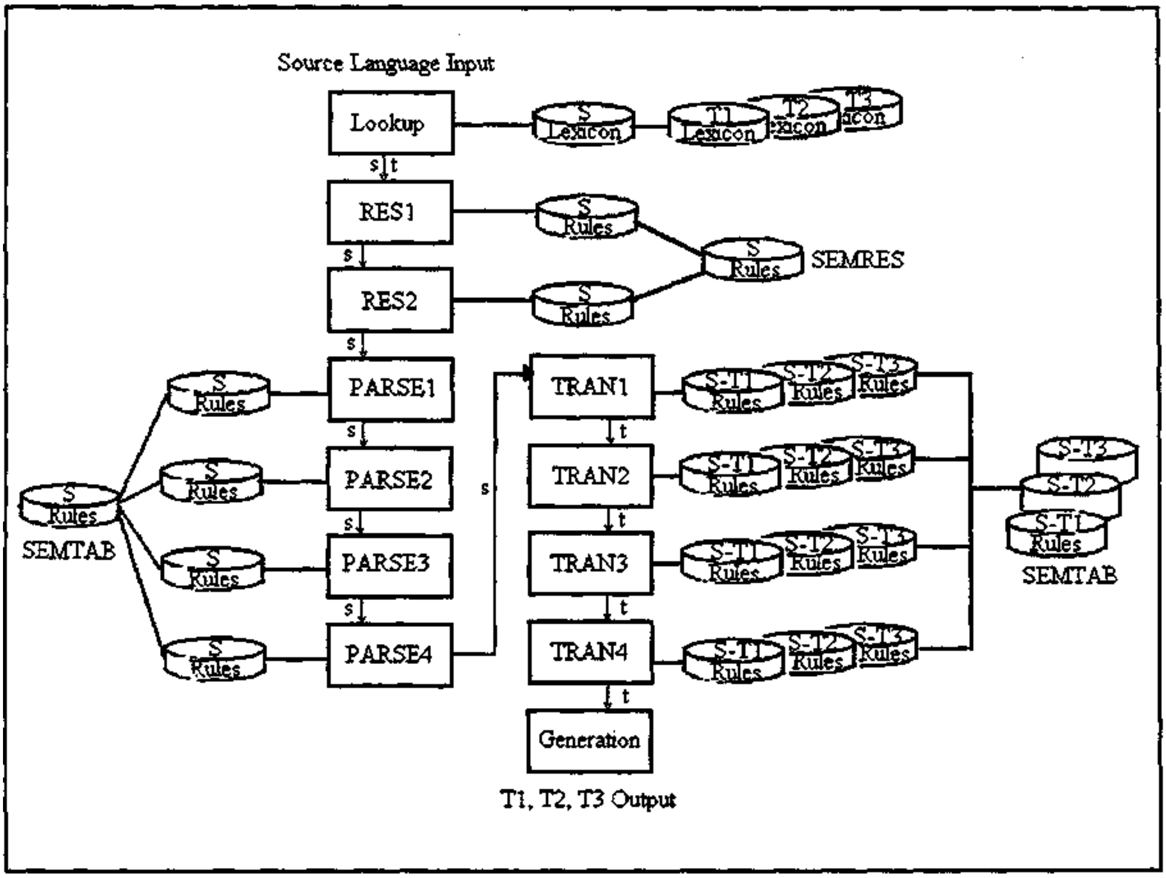

**Figure 5**

Once accurate information about the sentence's constituent structure is available, the problems of translating *any books* and *May 1st* can be more easily solved. It will also be possible for a TRAN linguist to create a "transfer tree", i.e., to rearrange the source tree so that extreme divergences from the target are minimized. Thus, a noun and a relative clause can be represented as one phrasal element for English-French transfer, while they can remain two separate elements for English-German. TRAN linguists even have the possibility of creating a transfer tree which

[6] Of course, linguists working on similar target languages will continue to cooperate in strategy development.

differs radically from the analysis tree. For example, the transfer tree for *A drop in temperature* I *occurs* can be made to look like the tree for *The temperature drops.*[7]

Since less time will be spent reconciling rules, linguists can devote more time to improving the system. Other advantages of the fourth-generation model for development methodology include:

- Target rules can be simplified. In the third-generation system a target rule must include all possibilities for target generation that are implied by a source pattern. In the new system, multiple patterns can be written in a transfer module, according to the source distinctions that need to be made for a given target. The result is shallower rules which are easier to maintain.

- Linguistic development becomes more stable and predictable. The PARSE modules will send analysis information to the TRANs in the form of an information array. The TRAN linguists can use this information as their priorities require it. They will not be forced to address every change in source analysis immediately, as before. With this independence, it will be easier to schedule and keep development deadlines.

- A separate analysis component makes it possible for a linguist to focus exclusively on improving the parse, without having to be immediately concerned about how each target will generate the translation. This will result in consistent improvements in analysis.

The technology for the fourth generation is the same as that of the third generation. Linguists, both in-house and at remote sites, will continue to work in individualized, linked environments, using minis to test new rules and strategies. The latest complete system will be available in a central location.

## 6. Conclusion

The history of Logos has been constrained by the requirement of maintaining a commercial system. The transition from one phase of development to the next must occur without disrupting the continuity of users' installations. It is not possible to rebuild the system from scratch on a regular basis. Nevertheless, the foundations of Logos have proven to be solid, with the system continually improving over several generations of development. We have illustrated the interaction between advances in system design and advances in development methodology. In reaction to linguistic deficiencies, the MT system underwent changes. The illustrations in this paper show that the system became larger and more complex. But it also became more modular. Modularity made it possible for specialists to work on different parts of the system. Development methodology has advanced to allow each specialist to develop and test rules in his/her own environment and to merge rules into the system in a rational way. The newest Logos model makes possible the separate development of analysis and transfer components, so that all targets get the benefit of one parse, and the source and each target can be developed independently. New source systems will be designed based on the fourth-generation model, in the hope that their development will avoid the linguistic and methodological problems of the earlier generations.

[7] This is an example of a constituent shift. For a discussion of constituent shifts see Gdaniec and Schmid (1995).

## References

Gdaniec, C. and P. Schmid. 1995. Constituent Shifts in the Logos English-German System. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation,* 311-318. Leuven: Centre for Computational Linguistics, Katholieke Universiteit Leuven.

Warwick, S. 1987. An Overview of Post-ALPAC Developments. In M. King (ed), *Machine Translation Today: The State of the Art,* 22-37. Edinburgh: Edinburgh University Press.

Scott, B.E. 1992. Biological Neural Net for Parsing Long, Complex Sentences. Internal Paper. Logos Corporation, Mt. Arlington, N.J.