Peter Wheeler

The LOGOS Translation System

LOGOS Corporation
111 Howard Blvd
Suite 100
Mt Arlington
NJ 07856-1307

LOGOS - Summary data

Name:	Logos
Status:	commercial
Type:	transfer with strong interlingual tendencies

Languages:	German-English	English-German
	-French	-French
		-Spanish

Speed:	installation-dependent

Cost: between DM 69 000 per year (VS, 2 user IDs)

and DM 315 000 per year (MVS, user IDs unrestricted)

Dictionaries: German source 90 000 entries
English source 50 000 entries

Data bases 13 500 semanto-syntactic rules
19 000 pure semantic rules

Implementation
language: Fortran 77

Operating system: CMS, MVS, Wang VS

Hardware: IBM mainframe (System/370 architecture including MVS/XA)
Wang VS

The theme of this conference, as it refers to the commercial systems on the machine translation market, is an overview of the changes which have occurred between January 1985 and the present. This is indeed a period in which Logos has made a number of far-reaching changes both in the corporation itself and in its product offering.

As 1985 began Logos was basically a one-product company, with three language pairs, only one of them at all mature, running on an obsolescent word-processor, (or worse still, in emulation mode).

As we enjoy the summer of 1986, all our Wang OIS customers are upgrading progressively to the Wang VS, i.e. from a word-processor to a computer, or to an IBM mainframe under CMS or MVS; we have added several significant new customers, among the best-known being perhaps Bosch, Deutsche Babcock, the Hoffman-Laroche Pharmaceuticals company in Geneva, and a little computer company known to film buffs as HAL …….. And to the rest of us as International Business Machines.

Having in the past concentrated our sales efforts in Europe, particularly Germany, we have now expanded our marketing effort to North America, with our sales force encountering very positive reactions there.

Development of English to Spanish is progressing apace, with this pair undergoing testing and evaluation at Burroughs; and development of German-French is well advanced.

We have moved our development center from Middletown New York, to Mount Arlington in northern New Jersey. (What we lost in picturesqueness, moving out of an abandoned railroad station redolent with a hundred years of steam and coal-dust, we more than recouped in such minor amenities as air-conditioning, a green-field site and a brand-new building.)

We have exchanged the uncertainties of long-distance time-sharing service for an in-house IBM 4341 installation.

On the commercial side, we have made major changes in pricing policy, changing from a pricing structure based on number of words translated to a flat-rate fee which has proved to be much more welcome to our customers.

We have established a regular pattern of biannual releases for all language pairs under development, and a totally common software for all pairs.

And all of this change and new direction is taking place against a background of uninterrupted development of the underlying Logos system, which makes it appropriate to start with a description of what Logos is and how it is used.

Logos is a menu-driven batch system, which in addition to Translation itself also incorporates Search features for Not-Found Words and Not-Found Phrases; and an interactive dictionary update facility known lovingly as Alex.

On certain systems, although not yet all, Alex has a sister Semantha, doing for semantics what Alex does for lexis. Plus a dictionary, of course, and the odd linguistic rule or two.

Let us look at the way the customer uses Logos.

Whether on Wang or on IBM, the user always starts from the main Logos menu.

OPTION:

Examine Input	Status Functions
1) New Word Search	4) Display TRD Status
	5) Clean Up Work Disk
	6) Print TRD Status
Translation	Utilities
2) Translate	7) Profile
3) Never EOS Utility	
	Dictionary Functions
	8) Add Entry (ALEX)
	9) Delete Entry
	10) Display Entry
	11) Dictionary Utilities
	X) EXIT

Fig. 1: Main Logos menu
He calls up "Translation":

Logos Translation System   G-E Rel. 3.0
Start a Translation   PANL1329

Enter Input Document: _____

Enter Output Document: _____

Enter Subject Matter codes for this Translation	Enter Company ID ATA
000 000 000 000 000	Select SMC Default X General Technical
Comment, if any: _____	Flag Unfound Words X No Yes

Fig. 2: Logos translation menu

-23-

and identifies the document to be translated, in the Wang environment by a word-processing document number, in the IBM mainframe environment by a file name and file type. He gives his company code, to ensure that the words he has himself put into the dictionary will be selected during translation in preference to anyone else's, and he gives one or more subject matter codes.

The Logos system offers 246 different Subject Matter Codes, (linked internally by a superordinate generic code level), and up to five of them can be given, in order of priority, to steer the choice of terminology towards a given subject field.

Let us suppose our translator wishes to translate the sentences:

- The boards are new but the drives are broken. They have three keys and five defaults.

Translating with the Subject Matter Code 000, for "General Usage" will give:

"Die Bretter sind neu aber die Antriebe sind gebrochen. Sie haben drei Schlüssel und fünf Versäumnisse" or "Les planches sont nouvelles mais les trajets sont cassés. Ils ont trois clefs et cinq défauts," respectively.

Translating with the subject Matter Code 224, for "Data processing," on the other hand, will yield

"Die Schalttafeln sind neu aber die Laufwerke sind gebrochen. Sie haben drei Tasten und fünf Standardwerte" or "Les cartes sont nouvelles mais les unités sont cassées. Ils ont trois touches et cinq valeurs par défaut."

After hitting the "Enter" button, the translator can then turn to other tasks, or even go home! The translation runs in background, so that while it is going on the translator's terminal is available to him for editing a previous job or preparing the next. Most of our customers, indeed, batch up their jobs for overnight running, so that arriving at his desk the following morning, the translator will find a number of draft translations, which he will then post-edit up to the desired standard - either initially on paper or direct on screen, depending on his own preference, using whatever word-processing program is available on his computer.

It has been our experience, as reported to us by our customers and as verified in trials of our own, that the properly trained Logos user can post-edit up to four times as fast as he could have translated the job himself.

Provided, of course, that all or most of the words are in the dictionary.

Particularly in the early stages of an installation, the new Logos customer will precede the translation stage by running his text through New Word Search. This provides a list of any words in the text but not in the dictionary, and also, in the case of German compound words, includes suggested translations for unknown compounds whose component parts are known.

```
AKRON                              1
ANGELO                             1
COLMAR                             1
EINFACHEXPANDIERENDEN              1
FENSTERSCHEIBE                     2
     FENSTER SCHEIBE
     WINDOW - DISK
FERTIGUNGSVERFAHREN                1
     FERTIGUNG VERFAHREN
     PRODUCTION - PROCEDURE
KEILFORM                           1
     KEIL FORM
     WEDGE - FORM
MIREVAL                            2
MONTPELLIER                        1
MOTORRENNBAHN                      1
     MOTOR RENNBAHN
     MOTOR - RACETRACK
REIFENPROFIL                       1
     REIFEN PROFIL
     TIRE - TREAD
SAN                                1
```

Fig 3: Logos New Word Search output list

In this case we see that the words "Reifenprofil," "Fertigungsverfahren," "Keilform," and "Fensterscheibe" are not in the dictionary but the various parts of the compounds: "Reifen, Profil, Fenster, Scheibe" etc. are. The New Word Search output document thus shows us the compounds broken down into their components, and offers us as a translation of the whole the translation of the parts. In the case of "Reifenprofil," or "Fertigungsverfahren," we will find the translation adequate, because, for example, the translation of "Reifenprofil" is indeed only a translation of "Reifen" together with a translation of "Profil." It is thus unnecessary to put "Reifenprofil" into the dictionary.

Even though it does not "know" the compound, the system will handle it correctly, giving it the combined translation of its component parts and giving the whole compound the syntactic and morphological characteristics of the second part, "Profil."

(With "Fertigungsverfahren," note the correct handling of the median 's.')

"Wedge form" as a translation of "Keilform" is not correct, however, "wedge shape" being preferred, while the offering of "window disk" - a translation of "Fenster" followed by a translation of "Scheibe" - is of course clearly wrong. In this case the translation of the whole is not simply equivalent to the sum of the translations of the parts, and we will have to add the compound "Fensterscheibe" to the dictionary as "window pane."

Note, too, that New Word Search is sensitive to subject matter. As with Translation itself, up to five can be given, in order of priority. In "Reifenprofil," for example, New Word Search has correctly offered us "tire tread," since the lexicographer specified Subject Matter Code 197 - Motor Vehicles. If "profil" had come up in a text running under Subject Matter Code 126 - Aeronautics, or 048 - Railroads, on the other hand, NWS would have offered us "wing section," or "load limit" respectively.

Our customer, then, concluding his New Word Search, has a list of unfound words in the form of a Wang document or an IBM file. Using normal Wang or IBM edit functions he discards from the list the words he does not intend to put into the dictionary (i.e. compounds correctly resolved like our "Reifenprofil," words where the number of occurrences recorded on the list is so low as to make inclusion pointless, and words which were unfound merely because they were misspelt in the input) and then writes into the same list the target language translations for the words which remain.

Now he can re-enter the Logos functions, via the main Logos menu:

<div align="center">Logos Translation System   G-E Rel. 3.0</div>
<div align="right">PANL0lGE</div>

OPTION:

Examine input   Status Functions

1) New Word Search   4) Display TRD Status

                                      5) Clean Up Work Disk

                                      6) Print TRD Status

Translation   Utilities

2) Translate   7) Profile

3) Never EOS Utility

                                     Dictionary Functions

                                    8) Add Entry (ALEX)

                                    9) Delete Entry

                                  10) Display Entry

                                  11) Dictionary Utilities

                                  X) EXIT

Fig 4: Main Logos menu

and call on the services of Alex. Alex presents him with a menu containing
the first of the source-language unfound words from the New Word Search list,
the target-language translation he has already given it, and a number of
questions to be answered. Difficult linguistic issues such as "Is this noun
masculine, feminine or neuter?" or "Is it singular only, plural only, or both?"

Logos Translation System   G-E Rel. 3.0
Dictionary Function: Alex   PANL01631

Enter German word in canonical form   Subject Matter code

Keilform_____   000 GENERAL USAGE

Part of Speech

X Noun	Entry type	Source gender	Source number
Adjective	X Word	Masculine	X Singular and plural
Adverb	Abbreviation/	X Feminine	Singular only
Verb	Acronym	Neuter	Plural only

Keilform_____

English target

wedge shape_____

Fig 5: Alex dictionary updating screen

This is followed by one of the most significant phases of the whole Alex
operation, shrinking the word down to its semantic head and thereby
identifying the semantics of the whole word.

Logos Translation System   G-E Rel. 3.0
Dictionary Function: Alex   PANL01632

Enter German word in canonical form   Subject Matter code

Keilform _____   000 GENERAL USAGE

Part of Speech

X Noun	Entry type	Source gender	Source number
Adjective	X Word	Masculine	X Singular and Plural
Adverb	Abbreviation/	X Feminine	Singular only
Verb	Acronym	Neuter	Plural only

If your source word is a compound word or a phrase, enter the head element

form_____

English target

W edge shape _____

Fig 6: Selecting the semantic head

In the case of "Keilform," for example, the lexicographer will indicate that the head is "Form," and will then be asked whether in this specific case "Form" is semantically a pattern or shape, a grammatical voice, a type or model, a cylinder, a mould, a section, the lines of a ship, or maybe even something else!

<div align="right">
Logos Translation System   G-E Rel. 3.0
Dictionary Function: Alex   PANL02380
</div>

Select meaning closest to meaning of Form in your word

Keilform

X pattern or shape, i.e., property of a thing

- voice, i.e., symbolic data

- type or model, i.e., classification of something

- tuyere or cylinder, i.e. conduit

- mold, i.e., receptacle

- section or block, i.e., portion of something

- lines of a ship, i.e. configuration

- none of the above

Fig. 7: Selecting semantic category

Why?

Imagine the sentence:

Durch die Keilform des Autos wird eine schnellere Beschleunigung erreicht.

If our lexicographer, even while entering the correct lexical equivalent for "Keilform" in the dictionary, had given it the wrong semantic classification, an incorrect translation would result. Suppose that it had been coded as a conduit, as the type of cylinder used in glassmaking or steelmaking. The preposition "durch" would then be resolved as governing a cylindrical object, and would be translated "through," giving "A faster acceleration is attained through the wedge shape of the car." Only by a correct classification of "Keilform" as some sort of shape, i.e. a characteristic or property, will "durch" be correctly transformed to give:

Because of the wedge shape of the car, a faster acceleration is attained.

With the word thus codified in the dictionary, the lexicographer will call up the next word and repeat the process. A couple of caveats are in order here. The first is that this step-by-step description has made the process sound lengthier than it really is. We have found that with proper training - and we always provide training with every new installation - the lexicographer can "Alex" words at a rate approaching one a minute.

The second is that the proportion of time spent by the customer on dictionary-building drops as the weeks pass. Broadly, we have found that after about six months the customer will have put into the dictionary virtually all the words which he needs, which we have found will amount to between 5000 and 10 000 per subject field. Consequently, the proportion of time devoted to lexicography drops progressively, and after this initial period the customer may no longer even bother to carry out New Word Search.

Transformations for specific words rather than for classes are handled by pure semantic rules, giving a real translation rather than the mere transcription we would get if Logos was simply a dictionary look-up system. For example, the translation of "make" will be different depending on whether it is a "mistake," a "cake," or a "statement" we are making. These distinctions are handled by Semtab (for "Semantic Table") rules, giving us "begehen" when the object is some form of error, "vorbereiten" with something edible, or transforming "make a statement" into "behaupten."

Further, these rules may take the form of linking together disparate words and giving each of them a different translation ("mit" + "freundlichen" + "Grüssen"" is not "with friendly greetings," of course, but "Yours faithfully;") or linking them together but giving a composite translation for the whole group ("in" + "Lösung" + "gehen" is not to be translated "go into solution" but as "dissolve"); or giving a different transfer to words as dependent on their context ("jüngst" with all information nouns such as "Buch, Bericht, Ausgabe," etc. is not to be translated "youngest" but "most recent.")

This module does not occur at the end of the translation process, but at any desired point in the multi-pass process.

There are some 19 000 such rules in the German-English system, 7600 in E-F and 8100 in E-G, the data-bases for E-Sp and G-F are of course under development.

It seems clear that as a system approaches maturity, the proportion of linguistic improvement to be achieved via such so-called Semtab rules increases at the expense of the purely syntactic ones. They are essentially the fine-tuning rules.

One of the features on our word-processor-based language systems was the facility for the user himself to add his own semantic rules. He would enter a menu offering various syntactic templates, (e.g. "Verb + Noun" - as in "Entscheidung treffen," "Noun + preposition" - as in "Auskunft über," etc.) respond to the prompts and questions on the screen, and his rule would be generated. This perhaps above all others is the feature which enables the customer to tailor his Logos translation system to his own needs. Consequently, it is a firm intention of the company to offer the Semantha feature on all our hardware environments, for all language pairs, just as soon as the programming can be completed.

The major change in Logos over the period under review has been a reinforced application of the multi-target principles of the Logos design.

It was Logos' aim to develop a truly multi-target system, one whose source analysis would be written independently of the needs of the target. Not only does this approach achieve a more perfect analysis, but it also permits the time required for development of a new target to be greatly reduced. And this has been borne out by our experience over the past eighteen months.

The first English pair, English-German, has been in development for nearly five years and is still being worked on. The second, English-French, saw an initial and successful installation within 8 months of start-up, and after an additional eight months is now virtually on a par with English-German. At the beginning of this year we undertook English-Spanish and within six months made the first installation. This English-Spanish system will reach parity with its sister target languages in even less time than it took for French.

How is this rapid building of new systems possible? Precisely because we are not building new pairs, but grafting new targets on to an existing source.

Conceptually, the Logos System has to be seen as a body of re-write rules, each consisting of a pattern side and an action side. The action side typically re-writes the pattern as a more abstract node in the parse tree, and also annotates it. Going from re-write rule to re-write rule, the system works its way through the sentence bottom up, left to right, in a number of iterations until analysis is complete.

A trivial example of a re-write rule would be: Adj N = NP

But the action side of the rule may also wish to do something at this point for the sake of the target as well. While it cannot generate target structures at this time, it can annotate what that structure provisionally should look like. The linguist may wish, for instance, to indicate the ordering sequence of these elements for a given target language. The Logos system permits the linguist to indicate these things in a target annex attached to the action part of the rule, so that the rule would look something like this for French:

   Adj N = NP /N Adj/

and like this for German:

   Adj N = NP /Adj N/.

The multi-target architecture of the Logos system allows a given rule to have any number of target annexes.

When source analysis is complete, the ball is passed to a generation module that will generate the target. This module consists of a number of so-called "generators", one for verb phrases, one for noun phrases, and separate generators for each type of clause. There is even a double clause (or sentence) generator as well for handling target transformations that bridge two clauses, so as to effect, for example, the subjunctive case in a that-clause when the verb of the main clause is of a certain type, or again, so as to effect the ordering of dependent clauses before main clauses.

These generators are written as purely target rules, independent of source analysis. As such they can be accessed by any source. We will return to this point later.

Much of what these target generators do is triggered by some annotation made by a target <u>annex</u> rule or by a rule in the Logos semantic table. For example, a source analysis rule designed to recognize the presence of the present progressive tense might have, in its German target annex, an instruction to the effect that the German target transformation required for this tense should be: "dabei sein + infinitive." The analysis rule with its annex would look like this:

   Be + Pres. Part. = VP(prp) /dabei sein + infinitive/

The French annex would contain:

   être en train de + infinitive/

Now when the target VP generator begins to process this verb, it responds to the instruction supplied earlier and effects the transformation desired.

From this it should be evident that when the Logos linguists set out to add a new target language to a given source - in this case English - what we have to do is to add the new target annexes to the rules wherever appropriate, and then to build these generator modules. While this is not a trivial task by any means, it is fairly straightforward. The gain of course is that the new target lives off the same set of re-write rules as any other target. As improvements are made to these analysis rules, all targets benefit simultaneously.

It should not be supposed - in case I have not made myself clear here - that the syntax of each target is assumed to be identical in every case. On the contrary, our multi-target approach allows the source sentence

I want the document to be checked for errors.

to be transformed in German into

Ich will, dass das Dokument auf Fehler geprüft wird,

but in French into the syntactically very different

Je veux qu'on recherche les erreurs éventuelles dans le document.

I want the man to be given a book

will yield

Ich will, dass dem Mann ein Buch gegeben wird.

and the syntactically very different

Je veux qu'on donne un livre à l'homme.

Future targets contemplated for English source are Japanese and Italian.

The ease with which new targets were added to the original English-German pair encouraged Logos in 1986 to add a second target - French - to the German source.

The German-French system will access the same French target generators as were developed for English-French. The German source analysis is more mature than the English source analysis, its data-bases for the dictionary, semantic table, and re-write rules are almost three times the size of the English source. Nevertheless, we anticipate having an equally mature German-French system after 14 months of development.

As a result, we now have on-going simultaneous development of French, German and Spanish targets off a common source, and simultaneous English and French target development off the German source. Moreover, all target modules are common to all source modules. Thus the French and Spanish target components are and will be common to both English and German sources.

We should be clear that what we are describing here is not yet the perfect MT model of pure and independent analysis followed by pure and independent generation. There is indeed in the Logos system an important transfer component that is pair-specific. The semantic table is still pair-specific, and so are, as we said earlier, these target annexes to the source analysis rules. The pair-specific transfer components, however, tend to leave the heavy work to the independent analysis and the generation modules, and to that extent it might be fairly said that the Logos system is a transfer architecture with strong interlingual tendencies.

This validating of the multi-target architecture, and the more recent breakthrough into multi-source, have undoubtedly been the most fundamental changes in Logos over the past 18 months.

But the more pedestrian work of fleshing out the various data bases must not be overlooked, crucial as this is to the quality of the translation.

Here, too, as much use as possible is made of the source-side work which has already been done. Our lexicographers use Alex just as the customer does to enter words into the dictionary, with the difference that when they are doing a multi-target conversion, part of the work is offered to them on a plate. Typing the word "table," for example, on the Alex screen for English-French will cause the boxes for syntactic and semantic information to be prefilled from the source side of the English-<u>German</u> dictionary. The lexicographer will thus see that the word he is to enter first is a noun, neuter, singular and plural, and that its semantics are those of a flat surface on which things rest. All he has to do is fill in the empty slots for "table" and "feminine." Hitting the keys again, he will find that "table" has appeared a second time, again a neuter noun, singular and plural, but that this time its semantics are for a collection of data. All that remains for the lexicographer to do in this case is fill in the empty slots for "tableau," "masculine."

In about a year and a half, 48 000 English-source entries were given French transfers in this way.

At the same time, the semantic rules have to be given a new target equivalent in a very similar way. During the initial development of English-Spanish, for example, from a start in January to a test installation in July 1986, 5 900 English-source Semtab rules were given a Spanish action component.

The same principles apply, of course, to the development of a German-French system on the basis of German-English. The data bases themselves are much larger, but the ways of converting them remain unchanged.

In addition to the multi-target work which has taken much of our attention, a major effort has been made during this period to extract as much linguistic information as possible from what initially appears to be no more than a format instruction.

Given a text portion like:

"The control unit allows the cylinder to be:
forced shut;
slid open;
released in the event of a power failure"

the parts of speech of "forced," "slid," "released," etc. were presenting us with the standard homograph resolution problems beloved of the English language. However, looking behind the words to the whole document as carried on any magnetic medium, we find that there are additional clues to help us. What appears on the printed page may be just these few phrases - what is on the tape or on the computer disc is much more:

> "The control unit allows the cylinder to be:
> #$%STARTLIST
> #$%LISTforced shut;
> #$%LISTslid open;
> #$%LISTreleased in the event of a power failure
> #$%ENDLIST"

Once we are able to use this information to tell the system that it is in a list mode, therefore, it can be allowed to look back beyond the colon on the first line, and conclude that "forced," "slid" and "released" must be past participles all governed by "be."

We have made exciting progress in extracting linguistic help out of this linguistic noise, and are confident of taking this new insight further in the future.

Other innovations over this time period have involved work to deal with photocomposition and formatting commands actually embedded within a text, so that they may be preserved in the output. A text may contain, for example, commands such as "Begin highlight (.bhl.)" and "End highlight (.ehl.)," and these might occur anywhere in a sentence:

> ...bhl. red pencil .ehl..

With English-German, the stratagem had been simply to attach the commands to the words and carry them through to the output, which works well enough for German target:

> ...bhl. roter Stift .ehl..

but not for French:

> ...crayon .ehl. .bhl. rouge....

What we have with a phenomenon like these formatting commands is a degree of linguistic noise right in the middle of a linguistic string, which must also somehow be translated. I am happy to say that good results have been achieved on this, giving:

> ...bhl. crayon rouge .ehl....

Getting the highlighting commands into the right place in the target looks like a relatively trivial task. Simultaneous multi-target and multi-source development looks like a bold breakthrough. But this combination of the broad conceptual strokes and the painstaking attention to detail has characterized Logos' development over the past eighteen months, and this is the way we intend to go forward to an even more successful future.