

Chart Translation

Martin Kay

Stanford University and
Xerox Palo Alto Research Center

Abstract

For efficiency reasons, Machine Translation systems are generally designed to eliminate ambiguities as early as possible even if delaying the decision would make a more informed choice possible. This paper takes the contrary view, arguing that essentially all choices should be deferred so that large numbers of competing translations will be produced in typical cases. Representing all the data structures in a suitable packed form, much as alternative structures are represented in a chart parser, makes this practicable.

1 Translation and Knowledge

Judging by the great increase in activity in machine translation in the last few years, an outside observer might easily conclude that researchers in the field had finally reached the goal of truly practical translation systems towards which they have been striving for some forty years. Maybe some breakthrough has occurred, or maybe all the little incremental efforts has finally proved just enough to push us over an important invisible line. Insiders know differently. If we somehow seem to be winning the game that has been going against us for so long, it is not because we have learnt to play better; it is because the goal posts have been moved. Simply stated, the market for low-quality machine translation has grown from nothing to one clearly worthy of commercial interest in a matter of two or three years.

Whatever the reasons may be for this change in the translation market, the Worldwide Web surely played an important part. Browsers now routinely offer happy explorers the opportunity to have the results of their quests translated into their own language at the click of a mouse. Their expectations of this process are, however, no greater than they were of the initial search. Web usage is essentially casual, even when a lot could turn on the outcome. If our search comes up with nothing, at least little time will have been lost. If we find something useful, it will be frosting on the cake—something we knew we had no real right to expect. So it is with a translation that is offered. If by reading it fast with little attention to detail, we seem to perceive something in it that touches on the subject of our quest, at best we may gain some useful information; at worst we will

be amused.

There will continue to be a market for this kind of translation until substantially better results can be produced with little or no increase in the price. But there is also a great and growing need for high-quality translation and, as I have argued repeatedly (Kay *et al.* 1994, Kay 1997), computers will do little to help fill this need until very large strides have been made towards building programs that could pass the Turing test—programs that could, in other words, successfully masquerade as human beings on the other end of a telephone or computer-mediated connection. My claim is that substantial proportion of the linguistic problems that need to be solved in order to achieve high quality translation are already fairly well in hand but that linguistics is a relatively minor part of what is required for translation. The remaining problems are not confined to any particular field of endeavor. Anything that a person might know, or believe, or suspect, or impute to the knowledge, belief or suspicions of another, could be crucial for translating the next sentence in some text. Furthermore, the sentences for which a good translation is possible only in the light of such nonlinguistic, unsystematic, knowledge are the rule rather than the exception. In short, translation is what is sometimes called an *AI-complete* problem.

My conclusion from this has been, and continues to be, not that computers are out of place in high-quality translation, but simply that they cannot be expected to do the job alone. The human contribution to the enterprise is indispensable. However, it absolutely need not take the form of actually making the translation and, indeed, substantial increases in the quality of current fully automatic systems may be possible with contributions by humans that know only a single language. A person that knew the source language and the subject matter of the material to be translated, might be called upon to answer questions about the meanings of particular words and phrases or the referents of pronouns and definite noun phrases. If the questions are chosen with care, and the answers interpreted at a sufficiently high level of abstraction, then they may contribute to translations into several different target languages. This is important in view of observation that, while most documents are not translated at all, those that are, are usually translated into several different languages.

Once beyond the narrow realm or meteorological reports, humans are always involved in translation when-

ever any importance attaches to the comprehensibility, not to mention the naturalness and fluidity, of the result. These people are called *post-editors* in MT jargon, *revisors* more traditionally. Their work could also be a source of input to a semi-automatic system for, if there are nonlinguistic problems that must be solved in the course of translating a document from English into French, then many of them can be expected to crop up again when the document is translated into German and, certainly, when it is translated into Italian. If the answers to even some of these problems could be inferred from the changes that the revisor makes, then the cost of the next translation might be substantially reduced. In the balance of this paper, I will sketch a framework for the design of translation systems that would allow for these, as well as a variety of other approaches to the problem of nonlinguistic knowledge in translation, to be experimented with and exploited in a broad and robust manner.

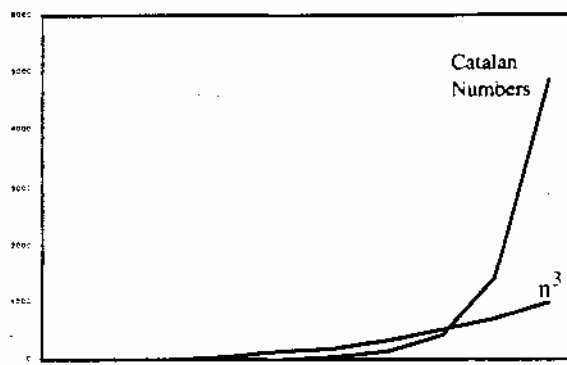
2 History

I begin with some history. A reasonable candidate for the year of birth of computational linguistics is 1960, the year in which John Cocke devised, without publishing, what later came to be known as the Cocke-Kasami-Younger (CKY) algorithm, the precursor of chart parsing. This is important for several reasons including the following:

1. It was the first algorithm designed to solve a major class of linguistic problems;
2. It separated program from data and grammaticality from considerations of contextual appropriateness;
3. It showed that an exponential number of objects could be computed and represented in polynomial time and space so that it was no longer necessary to curtail search paths prematurely.

Actually, the last of these was only partially understood at the time, when matters of computational complexity were generally much less well understood than they are today. More specifically, the finding was this: a representation of all the structures that an arbitrary context-free grammar assigns to a given n -word sentence can be constructed using time and space proportional to n^3 , even though the number of such structures grows in proportion to a series known as the *Catalan numbers*, which grows exponentially. As the following graph clearly shows, while the Catalan numbers initially grow more slowly than n^3 , the ninth

member of the series is 1430, almost twice $9^3=729$



This is a surprising result because, on the face of it, there is no way of doing anything whatsoever to a number, say k , of objects, and of recording the results for each of them, at a cost in time and space that grows more slowly than k . Lest any mystery remain as to how this comes about, consider the sentence (1) consisting of a noun phrase and a verb phrase, each of which contains an ambiguity resulting from a present participle.

(1) Visiting relatives would beat entertaining children

However, the ambiguities play no part in determining the well-formedness of the sentence as a whole and a single rule application is in fact sufficient to construct a representation of the four different structures that the sentence has. In other words, the chart parser constructs phrases, not out of other phrases, but out of equivalence classes of phrases belonging to a given grammatical category and covering a given part of the string. The structures that are built up in this way are referred to as *packed* structures.

The polynomial cost of chart parsing can be very high for long sentences, but the exponential cost of simple backtracking schemes places them entirely beyond practicality. Before the advent of chart parsing, it was therefore taken for granted that a syntactic analyzer could not be expected to deliver all the structures that a grammar assigned to a sentence but that it must attempt to find one, or perhaps a small number of, probably correct structures early and to abandon the search for any others. This reflects on my second reason for the historical importance of the CKY algorithm, namely that it allowed for a strong separation of program and data, grammar and algorithm. Such a separation is theoretically possible with backtracking algorithms, but impractical because such procedures had always to incorporate heuristics based on nongrammatical information to prune the search space early.

3 The Translation Relation

The proposal I want to make here is simply that the move that was made in parsing with the CKY algorithm in 1960 is the same move that needs to be made in translation. The ideas are based in large measure on ideas worked out in

some detail in Shemtov (1997). The results will be similar and the benefits even more dramatic. In particular:

1. This will provide the first complete implementation of the linguistic component of the translation process;
2. It will separate the translation relation from considerations of contextual appropriateness;
3. It will show that a very large number of candidate translations can be computed and represented at reasonable cost in time and space so that it will no longer be necessary to curtail search paths prematurely.

There is, of course, a potential problem with the parallel I am trying to draw between parsing and translation. Parsing has to do with grammaticality and grammatical structure. In other words, it is based on the notion of grammar as a set of rules that define a function from strings to sets of structures independently of external considerations such as the context in which the strings occur. The image of this in the translation domain will presumably be a set of rules associates strings in a pair of languages, possibly with structural information relating parts of one string with parts of the other, but independently of contextual considerations. The notion of grammar that parsers presuppose is an idealization that does not work for marginal cases and this will surely also be true of translation but, up until now, no one has seriously contemplated a context-free notion of translation, and it remains to be shown that such a notion is coherent

In this paper, there is space to little more than acknowledge the problem which, in its full generality is certainly difficult, subtle and multidimensional. Let us glance in passing at the well known problem of strings that can be translated into some other languages only by adding some information that is not explicit, at least locally, in the source text. Thus, we must supply pronouns with gender when translating from English into many other European languages, aspect when translating verbs into Russian, tenses when translating from Chinese, and so on. We must decide whether an English *chair* is a French *chaise* or a *fauteuil*, whether a *window* is a *fenêtre*, a *vitrine*, or a *guichet*, and whether a *book* is a *livre*, a *cahier*, or a *carnet*. In some cases, the choice is from a very large, possibly an open, set. Consider the French question *Où voulez-vous que je me mette?* which can be translated into English as *Where do you want me to X?*, where *X* can be replaced by any number of things, including *stand*, *sit*, *park*, *tie up (my boat)*, *sign (my name)*, and *leave my coat*.

For the moment, I will retreat to the position that cases like these are problems for the approach I am advocating to just the extent that they are problems for any competing approach. Any translation system is capable of rendering certain words and phrases in several alternative ways among which it chooses, sooner or later, based on more or less good grounds. At its weakest, the proposal I am making can be understood as nothing more than a recommendation that these alternatives all be left to be resolved by some

other module at a time still to be specified.

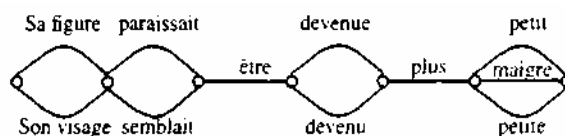
4 Charts and Contexted Sets

Consider the sentence (2).

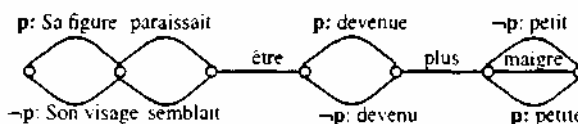
(2) Her face seemed to have become thinner
 Suppose, contrary to fact, that its only possible French translations are the following:

- Sa figure paraissait être devenue plus maigre*
- Sa figure paraissait être devenue plus petite*
- Sa figure semblait être devenue plus maigre*
- Sa figure semblait être devenue plus petite*
- Son visage paraissait être devenu plus maigre*
- Son visage paraissait être devenu plus petit*
- Son visage semblait être devenu plus maigre*
- Son visage semblait être devenu plus petit*

A chart can readily be built that comes tantalizingly close to representing this set. namely the following:



This chart could not, of course, have arisen as the result of parsing an input string because it has different word sequence covering one and the same stretch of the string. But the set of 24 strings that can be read out of it in the obvious way contain the 8 translations of the English sentence that we are interested in. The incorrect strings arise because there are dependencies among the edges that the chart mechanism itself is not sufficient to represent. In particular, the choice between *Sa figure* and *Son visage* determines the choice between *devenue* and *devenu* and between *petite* and *petit*. We can take care of this quite straightforwardly as follows:



All we have done is to annotate some of the edges with expressions from the propositional calculus—in this case, trivially simple ones. We now require that, when a string is read out of the chart, the conjunction of the expressions on the edges visited be satisfiable. This makes, for example

Sa figure paraissait être devenu plus maigre impossible because *Sa figure* can be part of the string only if **p**, and *devenue* only if **¬p**.

It is a short step from charts with logical annotations to the more general notion of *contexted sets*. A contexted set is a set of sets of arbitrary kinds of individuals represented in a particular way. Instead of representing it as a list of lists of individuals, we represent it as a single list on which an individual that belongs to any of the sets appears just once with an annotation in the form of a Boolean expression. Each different set of individuals corresponds to a different

assignment of the values **true** or **false** to the logical variables. This scheme is sufficiently general to accommodate representations for just about any imaginable abstract structure. Furthermore, under certain circumstances which occur routinely in linguistic structures, it has clear advantages over listing the members of each set separately. Intuitively stated, the circumstances are those in which the constraints on the sets to which a given individual belongs are independent of those governing most other individuals. This was not the case for the French translations of (2), but we do not expect interactions like these to involve more than small subsets of the words in a long sentence.

Consider the sentence in (3).

(3) Teachers like the English book.

I take it this has two different syntactic structures corresponding the paraphrases in (4) and (5).

(4) The English book pleases teachers.

(5) Teachers like the English ones make reservations

The following table contains deliberately simplified versions of its two structures in the form of a contexted set. Since we are representing structural information, we need the members of the set to be able to refer to one another and, for this, we use the number in the first column. The second column gives a part of speech and either a word or a sequence of numbers identifying other set members that are the constituents of the phrase. The last column gives the Boolean expression that determines which sets the individual belongs to. The logical variables are all of the form p_i , where i is an integer.

	Phrase	Context
1	N: Teachers	
2	Prep: like	p_1
3	Vt: like	$\neg p_1$
4	Det: the	
5	Adj: English	p_2
6	N: English	$\neg p_2$
7	Vi: book	p_3
8	N: book	$\neg p_3$
9	NP: <1>	
10	N: <5, 8>	$p_2 \wedge \neg p_3$
11	NP: <4, 10>	$p_2 \wedge \neg p_3$
12	VP: <3, 11>	$p_1 \wedge p_2 \wedge \neg p_3$
13	S: <9, 12>	$p_1 \wedge p_2 \wedge \neg p_3$
14	NP: <4, 6>	$\neg p_2$
15	PP: <2, 14>	$p_1 \wedge \neg p_2$
16	N: <1, 15>	$p_1 \wedge \neg p_2$
17	NP: <16,>	$p_1 \wedge \neg p_2$
18	VP: <7>	p_3
19	S: <17, 18>	$p_1 \wedge \neg p_2 \wedge p_3$

The sets for which p_1 is true contain *like* as a preposition. The complement of this set of sets contains *like* as a verb. The sets for which p_2 is true contain *English* as an adject-

ive. No set for which p_1 and p_2 are both true contains a sentence because, if *like* is taken to be a preposition and *English* is taken to be an adjective, then the string does not contain a sentence.

Let us now add just enough information to the table to suggest how alternative translations might be produced while continuing to profit from the compact representation. We will write French words in italics to keep them apart from the English ones.

	Phrase	Context
20	Det: <i>les</i>	q_1
21	Det: <i>des</i>	$\neg q_1$
22	N: <i>professeurs</i>	
23	NP: <20, 22>	q_1
24	NP: <21, 22>	$\neg q_1$
25	Vt: <i>aiment</i>	$\neg p_1 \wedge q_2$
26	Vt: <i>plaît</i>	$\neg p_1 \wedge \neg q_2$
27	Det: <i>le</i>	$\neg p_3$
28	Det: <i>les</i>	$\neg p_2$
29	Adj: <i>anglais</i>	p_2
30	N: <i>anglais</i>	$\neg p_2$
31	Vi: <i>réservent</i>	p_3
32	N: <i>livre</i>	$\neg p_3$
33	N: <29, 32>	$p_2 \wedge \neg p_3$
34	NP: <27, 33>	$p_2 \wedge \neg p_3$
35	VP: <25, 34>	$\neg p_1 \wedge q_2 \wedge p_2 \wedge \neg p_3$
36	Prep: <i>à</i>	$\neg p_1 \wedge \neg q_2$
37	PP: <36, 23>	$\neg p_1 \wedge \neg q_2 \wedge q_1$
38	PP: <36, 24>	$\neg p_1 \wedge \neg q_2 \wedge \neg q_1$
39	VP: <26, 37>	$\neg p_1 \wedge \neg q_2 \wedge q_1$
40	VP: <26, 38>	$\neg p_1 \wedge \neg q_2 \wedge \neg q_1$

We are assuming the English bare plural Teachers can be translated into French either as the definite *les professeurs* or the indefinite *des professeurs*. A new context variable, q_1 , is introduced to keep these apart. Each will be derived and represented only once, and then used repeatedly in translations like

Le livre anglais plaît aux professeurs

Le livre anglais plaît à des professeurs

Les professeurs aiment le livre anglais

Des professeurs aiment le livre anglais

Les professeurs comme les anglais réservent

Des professeurs comme les anglais réservent

The first two of these are the only ones we develop in the table. Needless to say, we gloss over the mechanisms required to reduce *à les* to *aux*, which clearly belong to a component of the system that we are not concerned with. The two noun phrases are items 23 and 24 and these contribute to a pair of prepositional phrases in 37 and 38. Together with the verb in 26, these give rise to the verb phrases in 39 and 40. The fact that these phrases differ only in the determiner in the object of the prepositional phrase is

reflected in the fact that the corresponding Boolean expressions differ only in that one has q_1 where the other has $\neg q_1$. Both have $\neg q_2$ showing that the verb will be *plaît* rather than *aiment*, and p_1 because the original English is construed as containing the verb *like* rather than the preposition.

This sketch has, of course, been very superficial. This is partly because of the obvious constraints of time and space, but it is also because I do not wish to give the impression that the approach to machine translation I am trying to advocate turns on using one particular set of algorithms or representing grammatical structures in a particular way, or adopting a particular approach to transfer. A more particular instantiation of a least some of the ideas is described in some detail in Frank (1999) in this volume.

5 Choosing the Best Translation

Our initial motivation for adopting something like contexted sets as the basic framework for a machine translation system had two components. On the one hand, we pointed out that it can introduce to a translation system the same kinds of efficiency gains, both in time and space, that charts make possible for parsing. If one wishes to pursue all paths through the search space, and if the chart constructed by the parser must be converted to disjunctive normal form before the next component can work on its output, then the claim that the parser operated in polynomial time will be of little interest. It is only if the advantages of the packed representation can be maintained throughout the entire system that the real advantages will be realized. Generalizing from charts to contexted sets makes this a quite reasonable goal.

The second advantage that we claimed for this approach comes from the sharp division that it makes between the fully automatic linguistic system that explores a large space of potential translations and a possibly only partially automatic component that chooses a single translation for each input. There are several reasons for claiming this as an advantage. The first is that of simplicity. Every machine translation system generates alternatives and chooses among them though few of them reflect the great difference between these two operations or the kinds of information that each of them calls for. Nothing but clarity can come from disengaging linguistic from nonlinguistic considerations in the way we are advocating.

Perhaps more important is the fact that the more traditional approach calls for choices to be made early, presumably on the grounds that this makes for efficiency. However, it also means that choices tend to be made among alternatives before they are completely specified and that the information on which the choice is made is partial at best. Idioms and technical phrases provide the simplest of examples. Suppose it is determined that the words *power amplifier* should be recognized as a set phrase on the grounds that its meaning is not a simple function of the meanings of the individual words and it is translated as a whole into some other languages. There is nothing wrong with this so

long as the possibility of its actually being two words on some occasions is not foreclosed. One such occasion would be the sentence *Amplifiers with higher power supply the main array of loudspeakers*.

The advantages of separating the generation of translation alternatives from the process of choosing among them go further. To simplify the discussion, let us assume that the set of alternative translations for a given string is so large that it typically contains at least one member—let us simplify again and say exactly one member—that would in fact be acceptable. That is to say that, after the generation of alternatives is complete, selection among existing alternatives is all that is required to complete the job. Within the framework we have outlined, selecting an alternative is tantamount to assigning truth values to logical variables so that the system of Boolean expressions in the contexted sets constitute a universal interface between the generation and the selection components. If a post-editor selects a particular string as the preferred translation into French, and the time comes to produce a translation of the same original into German, then those variables that occur in both the French and the German contexted sets will have the same values so that the choice that the German post-editor has to make will be among a smaller set of alternatives.

Suppose the sentence to be translated is *There are three windows in the room* and the post-editor chooses (7) rather than (6), then the choice in German between (8) and (9) is presumably determined to be (9).

- (6) Il y a trois fenêtres dans la salle
- (7) Il y a trois guichets dans la salle
- (8) Es gibt drei Fenster in dem Zimmer
- (9) Es gibt drei Schalter in dem Zimmer

Whether this would in fact happen, given the system as we have described it, turns on a technicality. If the word *window* is regarded as unambiguous, then the choice between *fenêtre* and *guichet* on the one hand, and *Fenster* and *Schalter* on the other, will presumably be made independently and there will be nothing to suggest the French choice correlates in any way with the German choice. But, if in the course of the English analysis that is common to the French and German systems, the English *window* is separated into *window*₁ (as in the window of a house) and *window*₂ (as in a ticket window) each of which is then translated unambiguously into each of the other languages, then the same logical variable will be associated with the choice and the post-editing of one will be reflected in the other.

6 References

- Frank, Anette (1999). "From Parallel Grammar Development towards Machine Translation—A Project Overview." This volume.
- Kay, Martin. Jean Mark Gawron and Peter Norvig (1994). "Verbmobil: A Translation System for Face-to-Face Dialog", CSLI.
- Kay, Martin (1997). "The Proper Place of Men and Machines in Language Translation". *Machine Translation*.

Maxwell, John T. and Ronald M. Kaplan (1991). "A Method for Disjunctive Constraint-Satisfaction" in Tomita, Masaru. *Current Issues in Parsing Technology*, Kluwer.

Shemtov, Hadar (1997). *Ambiguity Management in Natural Language Generation*. Ph. D. Dissertation, Stanford University.