[From: Manfred Klenner and Henriëtte Visser (eds.) *Computational linguistics for the new millennium: divergence or synergy?* Proceedings of the international symposium… 21-22 July 2000. Festschrift in honour of Peter Hellwig… (Frankfurt a.M: Lang, 2002)]

Practical Machine Translation and Linguistic Theory

Hubert Lehmann

linguatec Entwicklung & Services GmbH, Heidelberg
h.lehmann@linguatec-es.de

1 Introduction

Many of the roots of modern linguistics go back to attempts to translating the Bible into more or less every language in the world. But today translation is not really in the center of attention of most theoretical linguists. Communicating our experiences, thoughts and wishes in one language can be difficult enough, building a theory on how exactly we do this is even more difficult, and it looks like we are still only in the beginning of doing it, inspite of all the progress which has been made in the past hundred years. A theory of translation which deserves its name is even further off, since it not only requires the knowledge about the languages involved, but also how they can be put into relation to each other in a systematic way.

Yet attempts to translate automatically were made when the first computers came into existence in the late forties. The task was begun with a lot of optimism and gross misjudgment of its complexity. The spectacular failure was also the beginning of efforts to describe languages as formal systems, which was an absolute prerequisite for linguistics to become a science. We can say that machine translation was a very important stimulus for the linguistics we know today.

Now, fifty years later, where do we stand? Major questions of linguistics are still open, but machine translation is in fact used on a day-to-day basis. Some of the translations may turn out quite horrible, but many users find machine translation useful and practical. Have developers of machine translation systems found answers to linguistic problems which theoretical linguists are unaware of? Does theoretical linguistics need practical machine translation to make further progress?

If the idea is right at all to treat natural languages as formal systems, it must be clear today that these systems are highly complex and that it is basically impossible to test theories about them merely with pencil and paper. In my view - and I am sure I share this view with Peter Hellwig - natural language processing in the computer is the only way we have to really test linguistic theories. We both have used the computer to understand better how language works. But while he applied his results to teaching his students, I applied my results to building natural language systems which are helpful to users.

I will try to show in what follows how a machine translation system such as Personal Translator relates to linguistic theory, and, particularly, what theoretical questions arise from the development of such a system.

2 Implications from building a machine translation system

Translators often say that you need to thoroughly understand a text before you can translate it. You also need to know the intentions of the author, the purpose of the text for the new audience, etc. Speaking in linguistic terms, you need to have a full grasp of the lexical items, syntax, semantics and pragmatics of a text in order to be able to translate it successfully. To my knowledge, current theories are not able to model this complete situation to the extent that they could be applied to real-life texts. But when trying to develop a practical system we somehow need to come to grips with all the aspects of actual texts.

Ignoring the fact that there may be infinitely many sentences, we could do machine translation by storing pairs of sentences and their translations. Provided we have a quick way of finding a given sentence, and provided we have enough sentence pairs available, we may have an efficient system for doing translation. In fact, under the name of *translation memory system*, such programs have been on the market for about ten years, and they serve to dramatically increase the productivity of professional translators. If this is so, why bother building more sophisticated systems? The answer is that there are just too many sentence pairs to be stored, if such systems were to be applied to doing translation in general.

Hence to solve the problem in general, we do need a compositional approach to machine translation. Essentially this means that we need syntactic analysis and we need a method to map words and syntactic structures of one language to corresponding ones in another language. There is a multitude of approaches to these problems. For the sake of illustration let me mention one which became popular in the early nineties. I assume most of you know that at that time statistical approaches to speech recognition outperformed the analytical approaches. Therefore attempts were made to do machine translation based on statistics. The protocols of the Canadian Parliament which were available in both English and French were used as a corpus for training. And the method really worked - in principle. A number of sentences could indeed be translated with reasonable quality, but it took a lot of time and computing power to do the translation. And the amount of material needed for training such a system for practical use would have been immense, I suppose beyond the scope of what could be obtained and processed today.

Today's practical machine translation systems are all based on the analytical approach which is favored by linguists, even though theoretical linguistics has neglected a variety of issues which are essential for doing any natural language processing system. Language is about communication, it involves under-

standing linguistic utterances just as well as producing them. The focus of theoretical linguistics has been for a long time on syntax - from the point of view of generation - and to a lesser extent on the representation of the semantic content of utterances.

A machine translation system must be an integrated formalization of all the linguistic aspects of language understanding, production and translation. A corresponding theoretical model will have to account for all these aspects as well. So we need an integrated theory which can give a complete account of all these linguistic phenomena and processes. It is not just the question of how human beings acquire language, which preoccupied Chomsky, but for us it is much more important to understand how language is used in communication, and how the same content can be conveyed in different languages.

The issue of completeness is very important when it comes to practically usable natural language systems. When we started out with our work some years ago, we were surprised to learn that there was no complete account even of the morphological system of a single language. We had to invent suitable schemes of representation and we had to compile all the necessary data. Now, morphology is really one of the simplest issues in natural language processing, at least for Indoeuropean languages. The fact that nobody bothered to describe it completely and in a useful form, is symptomatic for the way linguistics has developed for many years.

3 The lexicon

The next important issue concerns the completeness of the lexicon. It has been one of the basic tenets of theoretical linguistics that the lexicon is finite. If this is so, why has nobody taken the time yet to record all the words of German? The assumption probably is unrealistic: new words are acquired every day and existing words become obsolete. Of course, a theory of syntax which abstracts from this fact may still be useful; but a theory of language must also account for changes in the lexical inventory, if it is to be realistic.

You are all familiar with the long and up to now inconclusive discussions about what is a word, or more precisely, a lexical unit in a language. From a formal point of view, the answer is simple: the lexical units are what you have put into the lexicon. Or, to put it differently: the lexical units are all those items you do not want to decompose into smaller units. The difficult part is to find the right criteria for distinguishing between lexical units and linguistic expressions which can and should be treated compositionally.

This issue is rendered more complex by the fact that a given linguistic expression may require a different treatment at different levels of linguistic analysis. Very often, expressions can be seen as syntactic constructions, but from a semantic point of view they are lexical units. Or they may be single words, but

still can easily be decomposed into smaller units which constitute a semantic complex. These observations imply that it is high time to find a reasonable and theoretically well-founded solution for the issue of what is a lexical unit, because this is the decisive prerequisite for building the lexicons of practical systems.

Leaving aside this unresolved issue, as system developers we have made our more or less well-founded decisions on what items to put into our lexicons, largely influenced of course by what traditional lexicographers have done before us. Now, we also must know what kinds of information should be associated with each lexical item. At least we must assign a part of speech. Yes, how many parts of speech are there, and how are they defined? Surely, this depends on the theory of syntax which is used, so we cannot decide this independently. But then we must make sure that really all lexical units can be assigned at least one part of speech. And the policy which says "if I don't know what it is, then it's an adverb" may not work too well in practice. For German, there are notorious cases where words are assigned more than one part of speech, because they do not fit very well into the conventional grammatical systems. Take *aus* for example as in

Der Herd ist aus.
Frieda macht den Herd aus.
Frieda nimmt den Braten aus dem Herd.

Is it really adequate to say that *aus* is an adjective, a verb particle and a preposition? Or take *außerhalb* as in

Hans wohnt außerhalb.
Hans wohnt außerhalb der Stadt.
Hans wohnt außerhalb von Berlin.

Does it really make sense to say that *außerhalb* is both an adverb and a preposition? If you agree with those grammarians who say that adverbs cannot have complements, then that is all you can do. In this case, I suggest you also try to analyze sentences like

Hans isst dreimal pro Tag Suppe.
Dreimal pro Tag isst Hans Suppe.
Hans isst dreimal Suppe pro Tag.

So again, with the assignment of parts of speech, we have a theoretically unresolved issue which admittedly depends very much on the theory of syntax which you embrace. Even more dependent on this are all the other types of syntactic information which should be associated with a lexical item the most important among which is the description of the complements which it can govern. Again, I mean a complete account for the lexical systems of the languages involved.

There have been attempts to achieve this kind of description for German as well as for English and French, but I am not sure how much attention to these efforts was paid by theoretical linguists. It is furthermore unclear how the syntactic system devised say in Longman's Dictionary of Contemporary English relates to the most current syntactic theories proposed elsewhere. It would certainly be interesting to see a comprehensive account of the English lexicon in terms of say LFG, HPSG, or GB.

As developers of machine translation systems we are dependent on theoretically well-founded information in the lexicon, and I think there are quite a few research tasks ahead of us inasmuch as the linguistic details of the overall lexicons are concerned.

4 Syntactic analysis

In current machine translation systems, syntactic analysis really is the crucial part. What type of grammar is needed in terms of the Chomsky hierarchy? In which way are the results of complexity theory relevant for parsing real text? What are the actual properties of concrete grammars used in existing systems?

In theory, the system we are using in Personal Translator is of type 0, i.e. equivalent to a Turing machine. For some theoreticians this might be the end of the discussion, because they feel that everything that is more powerful than a context-free grammar is unmanageable and therefore not interesting. My position here is the following: let us describe the grammars of the languages we are interested in in a form we feel comfortable with. If you are able to rewrite these grammars in context-free form, go ahead and do it. I think at this point in time, it is most important to arrive at a comprehensive formal description of the syntax of a given language at all. My main concerns at present are coverage and accuracy of the formalization, and there it is where we are looking for improvements. Once that has been done we can worry about complexity, theoretical elegance and other properties of interest.

I am not at all concerned at the moment whether our formal syntaxes have any resemblance to how we as human beings represent linguistic knowledge and how we use it to understand utterances. For me it is important to obtain responses from the program which are similar to responses of a person. When we can achieve this, we will know a lot about the linguistic competence of people, even if we have no direct access to the processes that actually go on inside.

Let me make one more observation about syntax: the syntax of a given natural language is a very complex and intricate system of rules, regardless of what approach we use. In my view it is simply impossible to test a syntactic theory with pencil and paper alone. I do not think there is any way you could prove in a formal sense that a given syntax is correct and complete. The only

way I see for validating a syntax or a syntactic theory is by trying to implement it on a computer and to test it against a large corpus of sentences.

5 Semantics and translation

How can one tell whether the linguistic analysis of a sentence was correct or not? One of the attractions of machine translation is that in many cases this is very easy to do. You just look at the translation which comes out. The thought may be unfamiliar to you, but try to regard the translation as a kind of semantic representation of the input. Surely, there is a big disadvantage involved: the translation is an expression of a natural language and not that of a disambiguated formal language. However, there are also advantages:

1 If the translation is wrong, then processing cannot have been right.
2 There is a non-arbitrary relationship between the constituents of source language expressions and target language expressions.

Compare this latter point to rendering natural language expressions in formulas of predicate calculus. How you relate natural language elements to elements of the formal language is completely arbitrary. So you could represent e.g. words as predicates or individual constants, grammatical relations as predicates or as arguments of predicates, etc. I am not saying that you should not use predicate calculus any more for semantic representation, rather I want to point out that looking at translation can also be quite revealing for the linguistic processing which goes on and also for understanding the linguistic structures of the expressions involved.

Earlier we said that for doing a good translation you first need to understand the text. Why is it not sufficient to do an accurate syntactic analysis and assign the proper translations to all lexical items, taking also into account the structural differences of the languages involved? The answer is clear: it is because of the context dependence of meaning. In order to reduce all the alternative analyses and choices of different translations for a given element, it may be necessary to understand the situation described in the text to be translated. Take a sentence like

The program was loaded.

Whether the translation is

Das Programm war geladen.

or

Das Programm wurde geladen.

requires knowledge on whether a state or a process is described. How can such knowledge be derived from the context of utterance? A linguistic theory which could solve this single problem would be very valuable for translating English or French into German.

A problem of many current approaches to semantics is that the purpose of a given semantic representation is not clearly defined. Machine translation can provide tasks like the one just described, and make the semantic analysis much more focused.

6 Differences between languages

There are many ways to describe a situation in a language. When you compare

John left the room.

and

John went outside.

you may very well describe the same situation with a slightly different perspective. When you translate between languages, you try to find pairs of utterances which describe the same situation in ways which are as close as possible. This is often very easy to do for language pairs such as German and English or German and French. In many cases words of the source language can be translated by words with the same part of speech of the target language. The same concepts exist in all these languages, because there have been shared cultural developments over many centuries.

From a theoretical point of view it is much more interesting to look at those expressions where conceptual and structural changes are involved in the translation. Inspite of all cultural parallelism, certain concepts, certain distinctions simply do not exist in some language. E. g., there is no good expression in English for the German word *Schadenfreude*, or there is a single German word *Schnecke* for the two English words *slug* and *snail*.

Structural differences come in many types. Take for instance

The pawn queened.

vs.

Der Bauer wurde in eine Dame verwandelt.

where an incorporated complement in English is made explicit in German and ergativity is rendered as passive. We have looked at many types of such relationships when developing our machine translation system. I am not, however, aware of systematic investigations of such types. I think they would be very valuable not only for machine translation but also for language learning.

7 Ambiguity

Irrespective of what level or stage of machine translation you are looking at, there is one prevalent theme, you could even call it the leitmotif of natural language processing - ambiguity.

All natural languages are ambiguous in various ways. In my view this is a necessity rather than an accident of language. As speakers of a language we need to adapt limited means of expression to an unlimited number of situations. As listeners we are experts in disambiguating utterances. For linguistic theory this means that we must try to understand the rules which govern the use and resolution of ambiguity. I think the role of ambiguity has long been underestimated, and this may be due to the fact that the processes which underlie the understanding of language received too little attention.

More than forty years ago, Yehoshua Bar-Hillel was the first to recognize the importance of ambiguity resolution for machine translation. He realized that one may need a rather deep understanding of the situation of utterance to accomplish this, and therefore he concluded that high-quality machine translation was impossible. The example he used as proof of his view was arguable given today's state of knowledge, but it is still easy to find examples which at least show the difficulties we are faced with. Consider a seemingly very simple sentence:

John moved.

with the following two translations into German.

John zog um.
John bewegte sich.

The translations show clearly that the original sentence is ambiguous. We can safely assume that 1. the ambiguity is real for native speakers of English, and 2. there are hardly any real-life situations where the ambiguity of this sentence cannot be immediately resolved by a hearer. But how can we give a formal description of when the first and the second readings are intended? Clues could be either in the linguistic context which preceded or which follows, or even in the extralinguistic situation.

I see it as a task for theoretical linguistics to provide the methods for making such formal descriptions, both for the linguistic context and the extralinguistic situation. A further task would be to come up with a typology of ambiguity which is more detailed than the distinction between lexical and structural ambiguity. I suggest to consider at least the various methods of ambiguity resolution to be part of such a typology. Take the example

They can the fish.

Here it is fairly easy to describe the criteria for ambiguity resolution in purely syntactical terms. For *can* to be a modal, you either need a bare infinitive which is governed by it, or it should not have a complement at all. This means the method of ambiguity resolution used here is based on the complements of the ambiguous word. Now look at the slightly modified example

They can fish.

In addition to the ambiguity of *can* between a modal and full verb reading, we are confronted here with the noun/verb ambiguity of *fish*, a type of ambiguity which is very common in English. There are several observations we can make:

1 Determiners like *the* and *a* often are very effective for disambiguating noun and verb readings.
2 The noun reading of *fish* is possible because it can be treated like a mass noun.
3 Even though we have two ambiguous words with at least two readings each in the example, we only get two readings for the whole sentence.

Now look at the even more reduced example

Can fish.

We do not get the modal reading of *can*, because we would need an imperative interpretation which is only possible for the full verb. But now we have an additional noun reading like in *sea fish.*

This sequence of examples has shown

1 how syntactic criteria can be used to resolve ambiguity
2 that there is an intricate interplay of such criteria
3 that syntactic criteria alone are not always sufficient to obtain just one reading for a sentence.

Semantic typing of lexical units and of slots for modifiers is a well-known method of ambiguity resolution which is widely used in machine translation systems. Inspite of many years of research and many attempts to build systems of semantic types, there still does not exist a system that most researchers would agree upon. We encounter again the problem of coming to an exhaustive solution. We should be able to assign semantic types to all of the vocabulary and to all the frames of modification. And we should keep the purposes of such characterizations in mind: in our case it is disambiguation for doing translation. So that we get the proper translations for e.g.

John grows a beard.
John grows sheep.

More methods are required to resolve ambiguities. In the end, we may be forced to model the situations of utterance to a large extent. But from an economical point of view, we would also like to know which techniques are most effective and how costly they are to develop and to use.

8 Conclusion

Currently machine translation is probably the most important field of computational linguistics. Machine translation has begun to become an indispensable tool for many users, and I assume that its significance will dramatically increase within the years to come. A prerequisite for this development is a steadily improved quality of translation, and this is what my company is working on with all the resources we can put into it.

So far the practice of machine translation has often preceded linguistic theory. This should change. I hope I have been able to point out some of the important areas where machine translation can stimulate research in theoretical linguistics. Quite a few of the issues I mentioned are well known to linguists, but others I believe so far have been ignored or at least neglected. In my view theoretical linguistics should take seriously the problems posed by machine translation and it should try to contribute as much as possible to its success. Machine translation cannot only pose new problems to theoretical linguistics or put known problems into a new perspective, it can also serve as an ideal testbed for linguistic theories.

I hope to share these views with Peter Hellwig who has worked on implementing linguistic theories for many years and has always tried to advance the state of the art also in terms of theoretical issues. And I would appreciate it if many of you would join in and contribute to the progress of both machine translation and linguistic theory.