BUDAPEST 1988:
Report on "New Directions in MT" and "Coling-88".

The first conference, "New Directions in Machine Translation", organized jointly by the Dutch firm BSO/Research and the NJSZT (John von Neuman Society for Computing Sciences) in Budapest, took place in the MTESZ building, in the centre of Pest, near the Houses of Parliament, on 18-19 August. It was attended by about 100 people.

The Conference opened with Professor Bálint Dömölki, President of NJSZT, who emphasized the need for minor languages to be translated into more widespread ones. Hence the particular importance of MT for them. Dutch and Hungarian were examples of 'minor' languages.

John Hutchins, of the University of East Anglia, and author of "Machine Translation, Past, Present and Future", gave a detailed overview of MT, and set the tone for presentations to follow. Indeed it was convenient for subsequent speakers to refer to Hutchins's expos as nothing had escaped his scholarly probing. The rest of the morning session was devoted to surveys of MT and MT systems. Hutchins expressed his doubts as to whether the prototype promised by EUROTRA for 1990 would be ready. The problems are twofold. One is that the coordination of activity in Europe, at first successful, now suffers from lack of agreement amongst the groups. The other is that, although the new framework looks more promising, getting the results is more complex.

He also pointed out that Prolog was becoming the most popular programming language for MT, not only for large projects, but also amongst individuals 'playing' with MT.

Dr. Ivan Oubine, of the USSR Centre for Translation of Scientific and Technical Literature and Documentation, gave a very down to earth account of the state of MT in the USSR. Ten years ago, he said, MT was criticized there for "burning the state's money". Today, the same people want to buy the same MT systems that were criticized. The reason he offered is that they are drowned under masses of information, in English and in Japanese in particular, and just want to know what it is all about. Quality of output is no longer an issue. The USSR has four MT systems which all only translate between a few language pairs (amongst English, German, French, Spanish and Russian). None is really multilingual yet. The speed of the various systems varies, so does the quality of output. Post-editing is needed. Curiously, poor outputs from English to Russian were criticized, whilst equally poor outputs from Japanese were deemed "very useful" or even "good quality". Oubine put this down to the fact that Japanese is not 'guessable'. In other words, any output from 'exotic languages' will always be appreciated, particularly as the aim is generally *assimilation* of information, (as opposed to *dissemination* of information, where top quality is required).

To help translators with terminology, they have an Automated Lexicographic system, which is a text-oriented dictionary, translating between English, German and French, with Russian as a pivot. Oubine also took the opportunity of his talk to advertise the fourth Moscow international seminar on MT, "Computer & Translation '89", organized by the USSR Centre for Translation of Scientific Technical Literature and Documentation, to take place in Moscow in December 1989.

Dr. Dong Zheng Dong, of the Language Engineering Laboratories in Beijing, gave an account of the development of MT in China. After the initial stage in 1957-65, it stood still during the period 1966-75 he said, mainly because of the bottleneck created by the processing of Chinese ideograms. 1976-82 saw a recovery, thanks to advance in grammatical analysis, achievements in semantics studies and, most importantly, a breakthrough in Chinese character information processing. The project he is working on, the Transtar system, is an Esperanto-based interlingual system for English and Chinese in the first instance. In his experience, Russian to Chinese translation is probably the easiest, whilst English to Chinese is the most difficult.

The afternoon session, chaired by Professor Brian Harris of the University of Ottawa, Canada, was devoted to interlingual architecture. Harris mentioned two Canadian MT systems, METEO and SMART. METEO, a transfer system, produces high quality output within the confines of its sublanguage, weather forecasts. SMART, an interlingual system, translates advertisements for jobs and produces 'bad' output. "Is there a connection between architecture and the quality of a system?" he asked.

Professor Christian Boitet, of GETA, University of Grenoble, has used a pivot language in the past and now uses a transfer system. He is dissatisfied with both. He saw the main drawbacks of a pure pivot approach as:
- the difficulties of concept lexical management (construction, maintenance and evolution);
- the limitations in MT output quality (mainly due to loss of lexical precision, giving paraphrase rather than translation).

In terms of cost, a pivot system may be cheaper than a transfer system but pivot lexicons are more difficult to construct. Also, a pivot system will only be cheaper for at least 8 languages.

Professor Iván Guzmán de Rojas, from La Paz, Bolivia, aroused a lot of curiosity with ATAMIRI, an interlingual MT system using the Aymara language, a language spoken around Lake Titicaca. This, he was proud to claim, was "the *smallest* project in the world!", in terms of people and money involved (the cost of developing the system was less than 1M dollars and there are at present 4 people on the team). Aymara is an old language and there are no technical words. The language representation is not based on networks but on matrices, with external tables for grammars, so that to introduce a new language, one simply uses a new table. At the lexical level, groupings of words are favoured (e.g. 'instead of' is treated as a cluster,

not as separate words). There is only an experimental prototype at the moment, which was demonstrated in Washington in 1985. There are translation centres in Holland, Germany and France, where the system can be seen in action.

Klaus Schubert, of BSO/Research, Utrecht, The Netherlands, rounded off the afternoon with an exposé on the architecture of BSO's own MT system: DLT, and asked himself "is it interlingual or double-direct?" DLT's interlingua, IL, is a modified form of Esperanto. The system performs two translations: from SL to IL, then from IL to TL: in that, it can be claimed to be double-direct (or even 'double-transfer'!), at least at the syntactic level. However, the semantic and pragmatic levels are dealt with by the knowledge bank and the word expert system, all written in IL: in this sense, the system is really 'interlingual'. This gave everybody food for thought...

On the evening of that first day, BSO/Research and MTESZ invited the Conference to a cocktail party in the very grand surroundings of the Ethnographic Museum. There, in the presence of her Excellency the Dutch Ambassador, the DLT project's Manager, Toon Witkam, thanked the Hungarian hosts.

The morning of the following day was devoted to discourse analysis and terminology. Dr. Christa Hauenschild, of the Technische Universität, Berlin, presented the case for the necessity of representing discourse structure in MT. This was being implemented in Kit-Fast (Kunstliche Intelligenz und Text verstehen — Functor Argument Structure for Translation). At the moment, the system deals with German and English only. Hauenschild gave the following example: "There was a bird flying in the sky. It was blue". The question was the interpretation of the anaphoric pronoun 'it'. Both 'bird' and 'sky' are possible antecedents, but 'a bird' is a newly introduced discourse subject, an indefinite NP (a definite NP would be less usual). One expects the sentence that follows to be about this newly introduced discourse subject.

Conference participants suggested that "There was a bird flying in the sky. It was blue and unclouded" and "There was a bird flying in the sky. It was blue that morning", would make the subject interpretation more problematic. However, it was pointed out that:

a. the two sentences above could be considered as 'bad English'
b. 'in the sky' is a locative phrase: it is unusual to refer to the locative phrase as subject in the next sentence.

Professor Jun-Ichi Tsujii, who has been working on the Mu project — amongst others — in Japan with Makoto Nagao, is to leave the Department of Electrical Engineering of Kyoto University to take up a chair of Computational Linguistics at UMIST in October 1988. He sees the discussions about different MT architectures as so many "wars amongst religions"! For him, current MT technology implies:

– a limitation in application fields;
– human interaction.

So the question is: how can we improve the quality? Tsujii believes that related research topics, e.g. dialogue translations, may help to pin-point problems.

He sees three types of interlingua, corresponding to three approaches:

a. interlingua as interpretation results, defined by formal concepts, e.g. physics.
b. interlingua as standard language, e.g. Esperanto in DLT.
c. interlingua as semantic primitives (e.g. in GETA), which are understandable by speakers of any language.

He believes that a. and c. will carry on to improve translations whilst b. is not so promising! Some expressions cannot be directly related to extra-linguistic facts. For instance, personal pronouns in Japanese will be translated differently according to whether the speaker knows the person in question.

e.g. I had to talk with **Mr. Smith**.

a. **he** is the lawyer who...: **sono hitto**
b. **he** is in Tokyo, isn't he? **kare**

There are currently some attempts at formalizing the pragmatics of discourse.

(A DLT project researcher was quick to point out that there were ways of inferring the information corresponding to Japanese pronouns in the DLT system!)

Dr. Christian Galinski, of Infoterm in Vienna, Austria, made a plea for terminology banks supporting knowlege-based MT. He sees the basic 'sins' in current terminology data banks development as:

– if it is conceptual only: it is soon too expensive.
– if it is linguistic only: it is soon useless.
– if it is only of one type (e.g. biology): then it has to be developed afresh for each other type (e.g social sciences).

This leads to lack of structure and coordination.

Galinski's claim is that this *must* be done by international cooperation — for instance by Infoterm! — as the job is too big for small firms or individuals. Pressed about costs, he quoted the example of 1 unit of terminology costing DM 5000! Admittedly, the cost included travelling and hotel expenses for the terminogra- phers...

The afternoon session, chaired by Klaus Schubert of BSO/Research, was devoted to Translation and Linguistics, and ended with an extremely entertaining presentation by Professor Claude Piron, of the University of Geneva, Switzerland, who claimed that "we can learn from mistakes made by professional translators". He introduced himself as both a translator (at the United Nations) and .. a psychologist. He expressed his pessimism as to whether MT could help human translators.

The day to day work of translators has got little to do with "blue birds", he remarked somewhat unkindly, and a lot more with things like: "Please mind the reception of all pickups problems", a notice stuck on the television set in his hotel bedroom in Budapest. Finding out what language it was originally in in order to retranslate requires *judgment*, he pointed out. How can a machine have judgement? Even in English, "to table a bill" is ambiguous. It can mean:

– to submit a bill to a legislative body (UK);
– to shelve, adjourn a bill (US).

English texts produced by Chinese for instance will either be in 'UK' or 'US' English.

Piron went on to claim that there was no such thing as a 'sublanguage'. For instance, the English word 'repression' can either be:
– French: 'répression', in a political context; or
– French: 'refoulement', in a psychological context.
In sentences such as "Repression by the population of its spontaneous critical reaction to...", it would be difficult to decide!

It was agreed that allowance has to be made for the fact that authors of English texts are not necessarily native speakers of English. But even in 'correct' native speaker English, problems can arise.

A translator, he claims, will translate "at 80 words a minute" and then will be suddenly stuck by a word like e.g. 'develop', which can have three different meanings:

1. set up, create: has no existence;
2. amplify, expand: has had an existence for some time;
3. tap the resources, exploit a potential: the potential exists.

If the text refers to Switzerland, there is no problem as ignorance can be quickly remedied (by a telephone call). But if it refers to — say — Tonga, the choice could be between 'set up' and 'expand' and will have to be sorted out before translation. It is these difficulties that take up, he claims, 80 to 90% of the time of translators, and not(i) finding out which is the subject, which is the object..

Naturally it was pointed out that his argumentation was a little unfair, that computers work in a *different* way. As Boitet put it, "planes don't fly like birds, but they fly". Also, the advantage of MT is consistency of translation, as large documents do not have to be split up between different translators. Oubine pointed out that at least, a computer will not invent when it doesn't know, but will stop (albeit too often, but at least, it is more honest!). Finally, it was pointed out that AI is not trying to do what humans cannot do, but do or at least understand what humans can do well and fast, and why.

On the evening of the second and last day of this Conference, the BSO/Research and MTESZ organisers exchanged presents to express their mutual gratitude for organising the conference. The BSO Conference was the first of its kind and participants expressed their wish to see a follow-up.

Saturday was an occasion to celebrate Constitution Day (or St. Stephen's Day for the reactionaries!) The whole town was festive, the sky was blue (but not the birds) and cameras clicked all day.

Then, just like in England, the weather changed suddenly, and it was a grey, wet and thoroughly miserable Monday morning that saw conference participants (over 650

this time!) plodding their way through puddles towards the Karl Marx University of Economics, further down the grey Blue Danube bank.

Coling-88 is the 12th of a series of international conferences on various aspects of computational linguistics. Most of the sessions — on semantics, formal models, understanding and knowledge representation, speech analysis and synthesis, discourse, software tools, computer assisted learning, parsing, syntax and morphemics, language generation, lexical issues ... and machine translation! — run in parallel. I attended sessions on machine translation.

To hold this conference in Hungary means that, because of COCOM restrictions on the export of computers, it was not possible to have comprehensive demonstrations of the systems. However, if Coling-88, like Coling-86 in Bonn, was largely dominated by the US, Japan and West Germany as far as papers given were concerned, it is also true that holding the Conference in an Eastern block country enabled a much bigger number of Eastern block participants, particularly students, to attend:

Hungary:	57
USSR :	27
Czechoslovakia :	23
East Germany:	19
Bulgaria:	14
Poland :	8
China :	5
Romania:	2

The Conference was opened by the ubiquitous Professor B lint Dmlki. The first session on MT started with Jun-Ichi Tsujii and "Dialogue translation vs text translation: an interpretation-based approach". Tsujii stated that both were difficult but presented *different* difficulties. In a dialogue (e.g. for hotel reservation), one must extract from input utterances the *important* information necessary for the generation of translation, by referring to the *aim* of the dialogues. In that, the problem is similar to that of other natural language understanding systems. Tsujii gave the following example of:

a. a Japanese sentence: "hoteru (*hotel*)-wa, tomodachi (*friends*)- to Disuko (*disco*)- ni ikitai (*to want to go*)-**node**, Roppongi (*Roppongi - the name of a district in Tokyo*)- no chikaku (*to be near*)- ga iino (*to be good*)- desuga."

b. a structure-bound translation: "as for hotel, because [I] would like to go to disco with friends, to be near Roppongi is good".

c. a less structure-bound translation: "I prefer to stay at a hotel near Roppongi, because I would like to go to disco with friends".

The 'important' parts are underlined. Tsujii claims that the rest (here the reasons), is peripheral, and does not have to be relayed by dialogue translation systems.

Professor Alan Melby, of Linguatech and Brigham Young University, Provo, Utah, gave the results of his testing of the BSO/Research's DLT system in a presentation entitled "Lexical transfer: between a source rock and a hard target": i.e. the translator is caught between the rock of the SL, which cannot be changed, and the hard target of all the possible TL outputs, of which one must be chosen.

To avoid any distortion of result caused by 'tuning' (i.e. putting a text through a system, correcting errors, putting it through again etc. until the result is near perfect), he adopted the word list approach. The test of the DLT system, which was carried out in 1987, consisted of a "secret text", reduced to its base form (the words) with some misleading words added (so the subject of the text could not be guessed). The words were combined and sorted alphabetically. Human translators then did a 'blind' dictionary update, i.e. they updated the dictionary for all possible senses and connotations of the words in the list, without seeing the text; the text was then given for translation, with all the words in the dictionary.

The test was carried out with about 800 words and consisted of four scientific texts from the EC. The results, he thought, were rather surprising. Examples of 'problem' words included:

- area: 'région' or 'partie'? This could not be determined by the context nor by the syntax.
- still: 'encore' or 'en outre'?
- wide: 'large' or 'vaste'?

The DLT team concluded that a lot more bilingual corpus analysis was needed.

Lexical transfer is undoubtedly difficult. Melby called for more international co-operation on building and using large bilingual text databases: these should contain at least 10 if not 100 million words!

He also proposed a contest for MT systems to be carried out at Coling-90 in Helsinki in two years' time. By then, some of the major projects should have a prototype working (e.g. DLT, EUROTRA). Melby expected between 3 and 10 participants.

Hiroyasu Nogami, of the Toshiba Corporation, Kawasaki, Japan, spoke on "Parsing with look-ahead in a real-time on-line translation system". The input is read until a conjunction of coordination is reached (e.g. 'and') or when certain requirements are fulfilled (e.g. a Sentence requires a verb, a Noun Phrase requires a noun etc.), or when structures with discontinuity are satisfied (e.g. as---as, the more---the more). This is done with an ATN with look-ahead conditions. Look-ahead parsing is claimed to be very effective for long and complicated sentences.

(It was pointed out that, in a number of cases, it would not be safe to reach conclusions before the sentence has finished. But if the sentence has finished, then it is not real-time translation! etc. etc. Boitet saved the day by pointing out that "he had seen the system in action" and insisted that "it worked".)

The Panel discussion, at the end of the first day, tackled "the real bottleneck of natural language processing".

Masaru Tomita (of Carnegie-Mellon University, Pittsburgh), kicked off by quoting, amidst roars of laughter, "linguistic sentences" as would-be examples of translation problems (they are, in fact, examples of parsing problems in the first instance):

e.g. "John hit Mary".
"Time flies like an arrow".
"The horse raced past the barn fell".
"The mouse the cat the dog chased ate died".
"John persuaded Mary to expect that he believes that she likes an apple."

and offered 'real sentences':

e.g. "All processes (programs) in the destroyed window (or icon) are killed (except *nohuped* processes; see *nohup(1)* in the HP-UX Reference); therefore, make sure you really wish to destroy a window or an icon before you perform this task."

"This window contains an HP-UX shell (either a Bourne shell or C-shell, depending on the value of the SHELL environment variable; for details, see the "Concepts" section of the "Using Commands" chapter)."

Tsujii then went on with a provocative "Why I don't care grammar formalisms" (sic) and their would-be 'elegance', as proposed by theoretical linguists, particularly syntacticians. But he does care about grammar formalisms for computational linguistics: e.g. ATN, ROBRA, GRADE etc.. The objectives of both are different so they must have different formalisms. "The grammar formalisms ignore mostly the processing issues. Linguists do not care processing issues in their formalisms just as we do not care grammar formalism!"

The whole discussion quickly turned into a battle of theoretical linguists vs language engineers.

Dietmar Rösner (Project Genesis, Darmstadt, FRG), pointed out the deficiencies of current linguistic theories for Natural Language Processing:
- mismatch between a problem discussed in theory and encountered in real data;
- failure to deal with non-grammaticality of input;
- dependence of linguistic theories on the English language. i.e. more grist to the Nagao/Tsujii mill...

An analogy was drawn between theoretical linguists and language engineers, and maths and civil engineers... and the language (civil) engineers were accused of wanting to do without linguistic theory (maths), with all the consequences this oversight might involve!

To the question "If you were to build an MT system, would you take on a translator?", Tomita answered: "Five years from now, yes, to improve the system. Right now, no, there are too many other problems to solve!!"

It was obvious that theoreticians and engineers were not going to see eye to eye on that day.

Makoto Nagao (Electrical Engineering, Kyoto University), chaired the Tuesday morning session, which started with Pierre Isabelle, of the Centre Canadien de Recherches sur l'Informatisation du Travail, Laval, Qubec, reporting on "CRITTER: a translation system for agricultural market reports".

CRITTER is a research-oriented, lexically driven transfer project, dealing with restricted sublanguages of English and French and is meant to be totally reversible. It already produces bidirectional translations. The monolingual dictionaries contain about 700 lexemes. The transfer component has about 300 correspondence rules. The grammars are still incomplete. It was pointed out that syntax is not that simple in cattle reports! The insistence on reversibility was queried, particularly when it had been shown not to work in some of the cases quoted. Isabelle said it was a means of guaranteeing the grammaticality of the output, since the input was assumed to be always grammatical. Somehow this failed to convince.

Wednesday was as usual a day off for Coling Conferences, reserved for tourist pursuits. On this occasion, the organisers took the participants by boat and under the rain to the Danube Bend to Esztergom (residence of the first Hungarian King in the 11th century) and back by coach (or vice versa: no boat would take 675 people!) via Szentendre, a picturesque old commercial town where a visit of a Serbian Orthodox Church and of the Museum of the ceramist Margit Kovács were scheduled.

The next plenary session was on "Trends and perspectives". W. Wahlster (University of Saarbrücken) saw computational linguistics as being 'theoretical' in the 70's and 'mathematical' in the 80's, unfortunately with a lot of definitions, few theorems, and no proof: in other words, not exactly good maths, hence the emergence of some "baroque formalisms".

Nagao saw 'a lot of action' since 1980 but no commercial packages that are any good for practical use. MT systems will be, if anything, "an *engineering* success" some time in the future!

All seemed to have something to say about the feuding schools of ideas, what with statements like "connectionism in its most religious form" and "formalisms are good for fighting and great fun to play with".

The last Panel discussion purported to discuss "The relation of lexicon and grammar in MT" or: when should compound expressions be entered whole in the dictionary

(as opposed to being parsed and analysed). Nagao summed up the discussion with the following earth-shattering declaration: "50% of compound nouns can be treated in a systematic way, 50% cannot..."

The otherwise lively and enthusiastic discussion prompted a (human) translator to exclaim: "After 40 years of translating, I greatly admire the optimism of some of you!"

Monique L'Huillier