History and Heritage of the DLT (Distributed Language Translation) project

Toon Witkam
Utrecht, Netherlands - 2006

I . Spark in the car

When I joined BSO (Buro voor Systeem-Ontwikkeling, Utrecht) in 1978, that software company, founded two years before, was still widely unknown. Some fifty people, most of them in their early thirties, were working there and it was my third employer after I had graduated from Delft Technical University as an aeronautical engineer. That was 8 years before. My graduation project had been about designing aircraft structures by means of interactive computer graphics, pretty much a pioneering piece of 'IT' at that time. So I set off for a career in the software industry. Working in Germany and the Netherlands, I was on a variety of computer projects dealing with science and engineering applications, eager to make good use of my math-minded technical education.

However, underneath there was also a feeling for language. In the good old days of the fifties and sixties, learning three foreign languages (English, French, German) throughout high school was still standard in Holland. On top of that, at the grammar school my parents sent me to, Latin and Greek were obligatory, and I also learned some Russian in my free time, fascinated by Sputnik, Gagarin and other pieces of soviet space exploration.

Esperanto I had never dealt with. It simply was one of those things that you knew existed, like the kangaroo in Australia or a planet called Pluto. I remember I had read somewhere that it was a 'perfectly regular' language, every noun ending in -o, every adjective in -a, and so on. So one day on my way to work, in the spring of 1979, I got a spark of inspiration. With my car radio tuned to the BBC World Service, hearing some talk about Honoré de Balzac, and being vaguely aware of the growing load of translations in the European Community, I suddenly imagined the perfectly regular Esperanto as the ideal intermediate stage between all the EC's translation pairs – the 'standard interface' so to speak in software engineering terms. I was excited and got hooked on this thing for months (and even years!) to come.

The first thing you do after such an 'invention' is to check whether it has not been invented already. So I began a hunt for literature on machine translation. Exploring the Esperanto world as well, I came across J.O. de Kat in Delft (who pointed me to Bruderer's *"Handbuch der maschinellen Sprachübersetzung"*) and the name of Heinz-Dieter Maas in Saarbrücken, each of them experimenting with computerized translation from or into this language. But nowhere did I find a definite plan or project in which Esperanto served as the intermediate language in a multilingual translation system, although such a role had been suggested for it once or twice.

The end of the seventies marked the getting together of two up till then fairly disparate fields: computers and telecommunications. With 'Viewdata' as a precursor, visions and implementations of international computer networks to exchange text documents began to arise. However there was a law, adhered to by all experts in the field, stating that transmission speed over communication links would always remain the bottleneck, certainly in comparison to the ever increasing processing speed of computers. This prompted system designers to go for compaction, and I was one of them. I had a hunch that Esperanto, with its very regular agglutinative structure, would lend itself to compaction quite well. To investigate this, I took on Gert-Jan Vlasveld, a graduating informatics student, for an internship in BSO. In a 6-month study, he found out that a variable-length (Huffman) coding scheme for Esperanto morphemes could produce an average of 19 bits per text word, instead of the 5-6 bytes usually needed with standard Ascii coding. In my view, this made Esperanto an attractive alternative to intermediate structures or 'languages' proposed by others [Andreyev; Booth; Droste; Goshawke] in the MT literature.

With the expanding videotex mass market and compaction-dependent transmission cost in mind, the first publication of what was later to become known as ‘DLT’ therefore took place at the Int’l Conference on New Systems and Services in Telecommunications, in Liège, November 1980. Similar settings at which I presented a paper were the IFIP Computer Network Conference in Budapest (May 1981) and the ASLIB EURIM5 at Versailles (May 1982). These first papers dealt mainly with the architecture of a ‘distributed’ translation system, as opposed to the mere batch MT. This would fit in the emerging large-scale networks (e.g. EURONET-DIANE at that time) with big information providers at the sending and a multilingual consumer mass market at the receiving end of the line. BSO would accordingly split the translation process into two stages: SL-IL and IL-TL. During text generation, the SL (Source Language) could be semi-automatically translated and encoded into the Esperanto-based IL (Intermediate Language), which would serve as the standard vehicle for text storage and transport. Computers at the other end then only needed to be equipped with a special decoder card for fast automatic translation into the desired TL (Target Language). What could be better?

II . Small is beautiful

Meanwhile at BSO my activities on ‘DLT’ had to take place largely outside daily working hours. In the evening of September 1, 1980, I had given my first in-house presentation of it, for 10 colleagues and Eckart Wintzen, BSO’s founder and president. Eckart had reacted with praise and enthusiasm, and had backed publication as well as patent application activity. Nevertheless, the company could not afford to keep me ‘unbillable’ for more than an occasional working day. Therefore, I asked for a 4-days-a-week contract, which gave me time to duly devote myself to my brainchild.

So I began studying textbooks about linguistics and its different schools by John Lyons, Geoffrey Sampson and others. On a more semantic track, I read Terry Winograd’s *“Understanding Natural Language”*, Yorick Wilks’ *“Grammar, meaning and the machine analysis of language”*, and Kelly and Stone’s *“Computer recognition of English word senses”*. I subscribed to the American Journal of Computational Linguistics and took notice of MT conferences and proceedings by Aslib. To counterbalance this Anglo-Saxon breeding ground, I worked up my knowledge of Esperanto. Despite all its regularity, it’s still a language you have to learn by practicing it and by reading novels (*“Sklavoj de Dio”* was to be my first). Preferring to take an objective stand on the language, I refrained from joining any Esperantist club or society, but I was eager to get in touch with esperantologists. At the nearby headquarters of the Universala Esperanto-Asocio in Rotterdam I got to know Victor Sadler, Simo Milojević and others, while in Hungary I met Dr. Antal Münnich, an inspiring scholar who at CLIDE ’71 had already proposed Esperanto for a role that came close to an intermediate MT language.

The beginning of the 1980s were marked by an upsurge in large-scale industrial innovation schemes. Japan challenged the Western world by launching its spectacular Fifth Generation Computers project. In response, Europe hastily set up the ESPRIT programme. Against this background, BSO was quite willing to seek public funding for innovative software such as DLT, and I worked hard to thoroughly describe and present its intricacies and possible development stages. A first and formal attempt to interest the Dutch government in subsidizing a 3-year pilot project failed. But then, in the summer of 1982, the big opportunity came from Brussels: the European Community Commission granted BSO 100.000 ECU (circa 250.000 guilders) to carry out a DLT Feasibility Study. For my employer and myself this was an enormous stimulus. The 1-year time frame for the study and its budget would enable me to work full-time on it, with the support of one in-house assistant and a couple of external advisors. In fact, it appeared a sound first phase of a possibly longer sequence.

My memories of the DLT Feasibility Study (Sept.’82 – Oct.’83) evoke the maxim “small is beautiful”. Its limited size allowed me to devote all my time to the content, hardly being distracted by organizational matters. As the non-linguistic elements of DLT (compaction, networked distribution) had been given enough attention in the few years before, I now could concentrate fully on the linguistic kernel, in particular the handling of hidden syntactic ambiguities, and how to represent the intended meanings unambiguously in the IL. As a full-time assistant I took on Alex Olde Kalter, an erudite graduate in Slavonic languages, who, besides working himself into Esperanto, helped me with extensive literature search and compilation. Examples of ambiguous constructions were given by various authors, but by far the most comprehensive source proved to be Erhard Agricola's *"Syntaktische Mehrdeutigkeit (Polysyntaktizität) bei der Analyse des Deutschen und des Englischen"* published in East Berlin in 1968. Perseverance in descriptive linguistics may be typical for Eastern Europe.

Anyhow, our EC-sponsored Feasibility Study had to rely on international input from at least a few EC countries. It was arranged that linguists at universities in four countries would support us with their expertise and constructive criticism throughout the year: professor Guy Bourquin and his team in Nancy, Dr. Heinz-Dieter Maas at Universität des Saarlandes, Dr. Bente Maegaard in Copenhagen, and Louis des Tombe together with Steven Krauwer from our home-town Utrecht University. They all were already involved in MT research, and discussing and checking the diverse syntactic disambiguations for DLT with them worked out well. As to the restrictions or modifications by which Esperanto could be turned into a totally unambiguous intermediary, additional consultation took place with Dr. Victor Sadler, Professor John Wells (author of *"Lingvistikaj aspektoj de Esperanto"*), and Professor Ivo Lapenna, all of them members of the Esperanto Academy. It should be noticed that there was no question of changing Esperanto itself: only a modified version for DLT's IL (Intermediate Language) was at stake, and the final responsibility for it was mine.

Ivo Lapenna, a prominent name in Esperanto circles, held professorships in International Law in London and previously in Zagreb. In connection with international business and economics, international law terminology was considered relevant for the text type to be handled by DLT. Lapenna's articles regularly appeared in *Internacia Jura Revuo*, and our interest was of course to consult him on law terminology in Esperanto. When he made his first visit to BSO, a small miracle occurred. Coming from abroad and driving his car to Utrecht, he took the wrong highway exit and landed in some nearby village. BSO was not exactly what you call a household name, but Lapenna did not mind: stopping his car along the sidewalk, he simply asked a passer-by "How do I get to BSO?", and was promptly given clear and correct instructions! As we later found out, the guy he had asked happened to be a BSO colleague on his day off.

In October '83 the DLT Feasibility Study was completed and presented to the European Commission, or rather to its 'Information Technology Task Force', in the form of a 350-page report. The report's poor matrix printer quality was made up for by a fair amount of carefully drawn diagrams and charts, and even some cartoons. On the cover an artist had drawn a modern-day version of the Dutch lion, proudly reposing after hard work, on top of a map of EC countries (10 at that time). Hard work it had been indeed: to save commuting time in the final weeks, I had been living on BSO's parking lot in my camper van.

We were rewarded! In the first place, Brussels immediately asked MT experts in the U.K., France and Germany (different from those who had been advising us) to study our report and to give their judgment. All three were clearly positive: *"The extensive presentation of Esperanto as an IL vehicle is most thorough and impressive"* (Frank Knowles, Birmingham); *"L'espéranto 'amélioré' est un choix judicieux. Le grammaire décrite dans le rapport correspond effectivement à un langage non ambigu."* (Gérard Sabah, Paris); *"In technologischer Hinsicht zukunftsweisend"* (Winfried Lenders, Bonn). There were points of criticism too: our treatment of each sentence in isolation (the neglect of text coherence), the chance of overloading the disambiguation dialogue at the SL side, and our predominant analysis of Esperanto at the cost of insufficient attention to both SL and TL. Nevertheless, the reviewers recommended the EC to give on-going support to the DLT pilot project, or at least to the parts proposed by BSO as the next step. Lenders added the advice that we should limit our text type to one specific field of application, instead of mere 'informative texts'.

Then there were the reviews in journals, partly praising, partly sceptical. Not surprisingly the choice of Esperanto raised doubts. Those who questioned the existence of up-to-date terminologies in that language, for a wide range of fields and disciplines, certainly had a point, though the reality was not as bad as some people presumed. A special category were the reviews in the Esperanto press. Our IL-directed disambiguation details were generally seen as sensible there; some of them were even suggested as potentially attractive for common Esperanto, e.g. the gender-symmetric pronouns we had worked out in the spirit of the 1980s (*li, hi, ŝi* and *ili, ihi, iŝi* for gender-neutral, masculine and feminine singular and plural

3rd person respectively). All in all, the main shortcomings of the DLT Feasibility Study, indicated by mainstream reviewers, were:

1. *"With a natural language like Esperanto as the representation language, it will be more difficult to facilitate inference making and deep understanding than with an interlingua that is based on a system of knowledge representation of the AI variety"* [Tucker 1984]. This observation came from the ranks of the more AI-minded MT researchers in America, in reaction to our claim that DLT's IL design was an 'excellent platform for AI enhancements', although no specific AI methodology had been mentioned in our report as an *essential* part of the initial design.
2. No clear description had been given of the separate SL-IL and IL-TL translation processes, and the tree transformations required for them. This view was expressed most articulately by [Bakel 1984] and also by [Neijt 1986], both of them linguists in The Netherlands.

Naturally, I got additional comments and feedback in later months, from meeting professionals in person. This included the ones that had been consulted by the EC Commission. So I met with Frank Knowles and later also with Siegfried Lenders, and BSO set up useful cooperation with each of them in the years to follow. But the important change in the immediate aftermath of the Feasibility Study was that I now felt part of the worldwide R&D community, when going to MT conferences such as Aslib in Cranfield 1984. It was as if I had passed my entrance examination. At the 1984 conference in Cranfield, over a cup of coffee, it was Systran's Ian Pigott who loosely uttered the words: "Maybe you can beat Eurotra".

III . The glamorous pilot

The promising outcome of the EC-sponsored Feasibility Study at the end of 1983 definitely opened the way for BSO to go ahead with a sizeable Pilot Project. As it happened, the follow-up sponsor was not the EC in Brussels but the Dutch Ministry of Economic Affairs in The Hague. An innovation-conscious wind was blowing through the Netherlands at that time, and The Hague came to an agreement with BSO's president Eckart Wintzen: over a six-year period, the ministry would subsidize the DLT project with a total of 8 million guilders (circa 3 million ECU), and BSO would invest an equal amount of its own.

Not only was the subsidy for a medium-sized (300 people) commercial software company a novelty, the long-term investment by such a company itself was also unheard of. Eckart explained his decision by pointing to language technology and AI as future growth areas in the software industry, and probably the innovative image that would radiate from engaging in this type of R&D will have appealed to him. Still, it remains Eckart Wintzen's great merit that he chose to carry out such an uncommon project as DLT within his BSO software house.

The 6-year DLT Pilot Project lasted from 1984 to 1990. An Utrecht-based team of 10 professionals, hired especially for this project, was involved full-time. In addition, a small army of temporary contributors was in place all the time, ranging from internships to occasionally contracted consultants or university staff at home and abroad. Project highlights were the Melby Test [see below] and our performances at COLING'86 (Bonn) and COLING'88 (Budapest). Apart from a pilot system demo of Simplified English to French via the Esperanto-based IL, some 1800 pages of R&D results were delivered in the form of published papers and a DLT book series.

Managing a project of this size and complexity often tends to become an awful job, but I was very fortunate in the circumstances under which I operated. There was mutual trust and respect between Eckart Wintzen and myself, and to accommodate the DLT Pilot Project a separate department, BSO/Research, was founded over which I was given complete control and responsibility. Twice a year I had to report progress to a committee appointed by the Ministry of Economic Affairs. We called this committee the "ABK": Bernard Al (Amsterdam), Harry Bunt (Tilburg) and Gerard Kempen (Nijmegen) were three university professors active in the linguistic or lexicographic domain, and a good working relationship was developed with them.

The most important thing at the beginning, of course, was to build up an excellent DLT team, with a good balance between linguistic know-how and software engineering. For the latter we got Crist-Jan Doedens, Bert Kessels, Dirk Mansvelder and later Ronald Vendelmans. On the other side, Bart Papegaaij, Job van Zuijlen and Dorine Tamis joined us as computational linguists. But the specific nature of DLT made linguists with fluency in Esperanto, such as Dr. Dan Maxwell, whom we later hired, indispensable. Right at the start I appointed two scholars who, in retrospect, were to become the main pillars of DLT and its heritage.

One was Dr. Klaus Schubert, an accomplished linguist with a fabulous command of languages. Having studied and worked at the universities of Uppsala, Kiel and Hamburg, this talented academic had published quite a variety of papers on grammar, computational linguistics and interlinguistics. He also had acquired fluency in Esperanto and cultivated a linguistic interest in it, regularly writing reviews in Esperanto journals. All this made him an ideal candidate for the post of DLT's chief grammarian. He entered the team at the start of 1985.

The other was Dr. Victor Sadler, whom I already knew from 1981-1983, when he had provided us with advice and cooperation. Sadler, a British national who had studied French and German language and literature in Cambridge and received a doctorate in London as early as 1962, had been using Esperanto as his primary working language for some 20 years. Over that period he had held editorial and managerial positions in the Esperanto movement, but in 1983 he had chosen to make a change and had gone to Alto Paraiso in Brazil, a remote rural area where he was teaching arithmetic at a primary school set up by the Esperanto foundation 'Bona Espero'. At the end of 1984 I wrote Victor a letter about the start of our big DLT Pilot, proposing a prominent role for him at the lexical side of the project's linguistics. We still lived in the pre-internet era, so when the time came to make final arrangements about the job offer, we did it over the telephone. When we had gone through the essential points and I began a chat, Victor interrupted me: "I have to hang up now, because I am standing here in the marketplace, and a whole bunch of people is gathering around me" (notice that it was also the pre-mobile-phone era). He joined BSO in the first half of 1985.

The Melby Test

Meanwhile, our new team member Bart Papegaaij had already begun work on what was to become SWESIL, the Semantic Word Expert System for DLT's IL (Esperanto). In the spring of 1984, we had received - via the scientific attaché at the Dutch embassy in Washington - an unexpected impulse from professor Alan Melby at Brigham Young University in Utah, in reaction to our Feasibility Study report. Alan Melby, who since 1970 had been participating in MT development of a multilingual system with some similarity to DLT, warned us not to make the error he had experienced: concentrating too much on grammar, and underestimating the problem of *lexical* ambiguities. DLT at its IL-TL side, with no post-editing, would be particularly vulnerable to such residual ambiguities, he argued. So, on his advice, we decided to plan towards a blind test which we called the Melby Test: our system would have to translate a text we would never see before, but for which we could prepare months ahead by receiving a list including all its distinct content words (but also other words, in such a way that we could hardly guess the subject field of the text). Judgment would be based on the translation of *lexical* elements alone, comparing these to the wordings chosen by competent professional translators.

So when Victor Sadler arrived and settled in Utrecht, he joined Bart and other team members to develop SWESIL, with the Melby Test on the horizon. Planning for this test was a perfect way of phasing a pilot project. It forced us to work as hard on the lexical as on the syntactic side, without the disproportionate investment that would have been required for covering a complete lexicon. Looking back, our exercises in word sense computation over those first years [Papegaaij, 1986], can be considered as a run up to Sadler's brilliant work in the second half of the pilot project.

The Melby Test with all its preparatory stages was a lengthy procedure, involving a sequence of moves between Utrecht and Provo (Utah), with quite a few people active on both sides. A large bilingual corpus on a variety of subjects from the EC and the UN was used to extract test material from. The final test, which took place in January 1987, comprised the SL-IL-TL translation into French of four English texts with 600 different content words,. Melby [1986, 1988] eloquently reported twice on the value of the test and its application to BSO's pilot project: first at COLING'86, when we were still in a preparation stage, and later at COLING'88, when he explained our reasonably passing of the test. The titles he chose for these papers are pretty revealing: *"Lexical Transfer: A Missing Element in Linguistic Theories"* and *"Lexical Transfer: Between a Source Rock and a Hard Target"*. We had something to be proud of again.

Real or unreal AI ?

Tucker [1988] in a review of our "Word Expert Semantics" [Papegaaij, 1986] showed himself impressed with the progress we had made in the three years since the Feasibility Study. Still, Tucker missed in the DLT model what he now referred to as *"real AI techniques",* whereas in his earlier review [Tucker 1984, cited above in Section II] he had referred to *"knowledge representation of the AI variety"*. In hindsight, Tucker's wording 'real AI' reminds me of what Doug Hofstadter [1979] once called Tesler's Theorem: "AI is whatever hasn't been done yet".

The truth is that in developing SWESIL, regular use was made of extra-linguistic elements and methods then generally considered to be part of AI: taxonomies, knowledge inference facilities, heuristic search techniques, probabilistic mechanisms. At the end of 1986, the size of SWESIL was 30.000 lines of PROLOG code, including a so-called blackboard model implemented for us by Jaap Noordzij [1987]. The chief software engineer in our team, Bert Kessels, became so involved in all this that BSO later formed a daughter company BSO/AI around him. Noteworthy is also the interest shown by Kessels in parallel processing and supercomputers at that time.

However, as the 6-year DLT Pilot Project proceeded into its second half, it became more and more evident that *"knowledge representation of the AI variety"* was not quite our solution to the MT problem. Instead, we chose the corpus-based approach: large amounts of text, in itself constituting a huge word net over which a powerful *semantic distance* function could be defined. Once compared to John Searle's *Chinese Room* paradigm, this was the excellent pioneering work of Victor Sadler [see Section V], and it was built upon Klaus Schubert's syntactic layer of similar quality.

Tesnière, not Chomsky

At a time when that method was still largely ignored in MT circles, DLT's chief grammarian Klaus Schubert decided Tesnière's dependency syntax was the most convenient method for a multi-language translation system. Klaus published extensively about this choice [Schubert 1986, 1987]. He argued that while constituency grammar is good enough for English, whose syntactical structure is based mainly on word order (constituency is in fact a word sequence), it is less useful for languages that are mainly morphology-based. When dealing with a diversity of languages, dependency syntax approximates to contrastive syntax (named 'metataxis' in honor of Tesnière), thereby facilitating the projection of structures from SL to TL, or SL-IL and IL-TL in the case of the interlingual DLT system. Schubert not only described and explained the principles of metataxis, but from 1986 to 1989 also organized the writing of concrete dependency syntaxes in 10 languages, including Bangla, Finnish and Japanese. All linguists involved were native speakers, and the results of their work were published [Maxwell 1989].

So under Schubert's guidance, the syntactic side of the DLT translation process with all the details of tree structures and tree transformations received full attention. In addition, three other points of criticism from 1984 were addressed. One was our treatment of each sentence in isolation. With regard to this, we published our "Text Coherence in Translation" study [Papegaaij 1988]. Another point was the advice to limit our text type to one specific field of application. Without giving up our attention to 'informative texts' in general, a substantial part of our pilot project dealt with AECMA (Association of European Aerospace Manufacturers) Simplified English, as used in maintenance manuals and technical documentation. On this, we were in touch with the nearby Fokker aircraft factories at that time.

Finally, our initially predominant concern to analyze Esperanto and ban out all potential ambiguities by minor changes to that language or at least its IL version, gradually faded. Towards the end of the pilot project, practically all of the modifications proposed by myself during the DLT Feasibility Study were abandoned. After all, contexts and corpora were now considered a much more important and durable

knowledge source for automatic disambiguation, in Esperanto too. And by keeping the IL equal to Esperanto, one could profit from the existing base of Esperanto texts, including manual translations, which made up for a (still modest but expandable) set of bilingual corpora.
At the end of the DLT pilot project I tended to conclude that Esperanto was in fact not the key issue in our technology any more. As an IL it remains interesting, even attractive, but the heritage of the DLT research effort [see Section V] is now marked by another jewel.

Drive and spirit

Within the Netherlands, our hosting company BSO enjoyed being in the spotlight during the 1980s as the model of an enlightened software house, and our DLT research project was part of that positive publicity, on TV too. Though the reporting about our goals and activities was not always precise, being in the spotlight helped to keep ourselves under a healthy pressure in our daily work, and to take good care of our contacts network.
In our own country there was hardly a university with which we did not have some form of cooperation: short contracts, consultancy, internships. From the many persons who contributed I must mention a few: Bieke van der Korst, Willem Meijs, Gert van der Steen en Hans van Keulen, all from the University of Amsterdam. Especially noteworthy is the Meijby test (named after Willem Meijs, as a follow-up to the Melby test) and the expertise on corpus linguistics which they shared with us. On the same subject we were in touch with Professor Frank Knowles in Birmingham, which was a center of innovative corpus work.

Our in-house DLT team worked hard and with concentration, but there was a spirit of openness regarding information exchange within the worldwide MT research community, which was estimated at some 1500 people then. Apart from giving demos and papers at COLING conferences, we organized a successful separate MT conference “New Directions in Machine Translation” [Maxwell, 1988] in Budapest. As to contacts with the Far East, I went to Japan several times, and Professor Makoto Nagao even honored us with a visit to Utrecht. The Japanese research environment somehow fascinated me and at the end of the DLT pilot project I spent two months as a guest researcher at the ATR Interpreting Telephony Research Labs near Kyoto. From Beijing we once hosted Professor Liu Zhuo and computational linguists Dong Zhen Dong and Li Wei for a three-day session at our office.
At our home base Utrecht we also had foreign guest researchers for many months, such as esperantist Ian Fantom from England and translator Claude Bédard from Montreal. In Ottawa I later went to visit Professor Brian Harris, the well-known authority on machine and human translation. Klaus Schubert in particular had a rich contact network, including celebrities such as Igor Mel’chuk, and we were also in touch with Esperanto linguists in Eastern Europe. In Zagreb we arranged with a local group to translate The Gaia Peace Atlas, in an effort to build up a bilingual English-Esperanto corpus.
Our contacts in the USA, except for Alan Melby in Provo/Utah, did not happen to be very fruitful. The BSO board had generously arranged a knowledge exchange contract with MIT for us, but in Boston we heard that the scientists there did not grasp why they were to be consulted by the Boston Symphony Orchestra!

IV . The missed production phase

During the second half of the DLT Pilot Project, planning ahead for a subsequent Production Phase began to take shape. The ingenious key methods for syntactic and lexical transfer via bilingual corpora [see section V] getting in place, it became clear by then that the next phase would be more of a challenge in logistics than one in technology.

Although Esperanto was not the crux of our corpus-based technology any more in 1989, it remained an attractive choice for the pivot role in a multilingual system, due to its structural transparency (including the formation of compound words) and its lack of idioms. On the other hand, esperantists available full-time with excellent translator’s experience and familiar with the terminology of a specific domain (e.g. economic news reports, IBM software manuals, etc) were scarce, and we needed at least 8 of them for the preparation and manual translation of a specimen corpus. The cost of making an initial bilingual corpus (with a modest size of 350.000 words per language) was estimated at 1.7 million ECU, and was recurrent for each particular field of application or text type.

In our different versions of a Production Phase plan for the 1990s, the total cost of setting up the production center and developing a product ranged from 18 million ECU for controlled language translation betweeen 4 languages to over 200 million ECU for unrestricted translation between 12 languages, including Chinese and Japanese.

Whereas cost estimation is always a tricky business, one thing was clear to me: we would need a team of people working together in one centre - a dispersed workforce would not do. Another thing was my conviction that a high quality MT system is never finished: it will need perpetual maintenance, to support new words and expressions in the languages, and to verify that the system's text corpora are updated in a well-balanced and reliable way. To increase the chance of getting the right people for the job, we had the idea that we might set up a DLT production and maintenance center in Central or Eastern Europe. In Eastern Germany, Poland, Czechoslovakia, Hungary and Yugoslavia there were at least more capable and experienced esperantists than in Western Europe, and we had a good contact network with those already.

Meanwhile BSO had hired McKinsey&Company to scope the market prospects for DLT. Armed with the outcome of this, we began (in 1989) to approach big players: worldwide publishers, IT and electronics companies. As an exercise in and showpiece of new technology the DLT research project had been great, but as a product it was completely outside the range of BSO alone, which was in fact a service-oriented, not a product-oriented company. In search of a partnership, BSO's president Eckart Wintzen went to Olivetti in Italy and to Robert Maxwell in England. A visit by myself to Rupert Murdoch's 2nd man in New York did not bring a result either, nor did an prolonged exchange of ideas and some cooperation with DEC in Southern France. In the USA, business consultant Gerry Haller arranged meetings for us with various assumed prospects. And some readers may remember the double-page advertisement campaign we ran in the magazine *'Language Technology'* through 1988-1989, with headings like: "Mr. De Benedetti, here's technology to open up Europe!" So we did try!

In 1990, the DLT Pilot Project came to an end. Up to 6 million ECU had been spent on it, about half of it paid for by BSO itself. As a follow-up in the form of an industrial partnership with a big investor did not materialize, this seemed to be the end of the DLT project as a whole. After another year, during which the remaining team carried out interesting MT contract work for IBM's Scientific Centre in Madrid, the former DLT team dispersed.

V . The treasure left behind

Towards the end of the 1980s, a paradigm shift in MT began to take place. The first clear signals were at COLING'88, where the IBM Mercer-group surprised the conference with the premiere presentation (by Peter Brown) of its Statistical MT, and where Junichi Tsujii made a courageous and persuasive attack on computational linguists spending their research time typically on toy grammars and sentences of the 'John-loves-Mary' type.

The paradigm shift, which took till far in the 1990s to be implemented in the worldwide MT research community on a large scale and gradually changed its MT-acronym into <u>S</u>MT, was characterized by <u>S</u>tatistical processing of bilingual corpora, as well as a dramatic drop in attention to grammar formalisms and linguistic theories. Perceived as huge archives of high-quality human translations, bilingual corpora were even considered to replace bilingual dictionaries as well, as the latter were only a limited and usually outdated by-product of the former.

In 1989 I witnessed the Paradigm Shift and contributed to it myself in Garmisch-Partenkirchen, at an MT seminar organized by the IBM Europe Institute [1989]. In a presentation, I argued that BITEXT [Harris 1988] and well-aligned bilingual corpora were the salvation for MT, and that dictionaries could be thrown away. For this viewpoint I was awarded the 2nd prize for the seminar's most 'heretical' paper. The first prize went to IBM's Robert Mercer.

My heretical plea however was nothing less than a reflection of the state of research in the DLT project at that time. Building upon his experience with SWESIL prototypes, and profiting from Klaus Schubert's silver layer of dependency syntax, our chief semanticist Victor Sadler had worked out a powerful and innovative scheme: the Bilingual Knowledge Bank (BKB), which was explained and documented in detail in the fifth book of our DLT series: "Working with Analogical Semantics" [Sadler 1989]. As this publication reveals, we were fully on the SMT track as early as 1989. We also based on it our plans and cost estimates for the production phase [see section IV], from which one may conclude that we did not underestimate the amount of work required for collecting and preprocessing large bilingual corpora, and putting in place the partial parsing and multi-level alignment software required for them.

A key element in Sadler's design was the alignment of so-called Translation Units (which do not necessarily correspond to syntactic subtrees, but may result from tree subtraction), in addition to phrase and sentence alignment. Another key element was the way in which the *semantic proximity* between any two words can be calculated from the BKB. For instance, if you have *noun1* and *noun2*, the dependency-parsed corpus (even if this is only partially parsed) will speedily produce an internal list of all the *Verb-Subject, Verb-Object, Preposition-Argument etc* dependencies in which *noun1* or *noun2* occurs, and the amount of overlap between the two determines their semantic proximity. So this measurement is of a statistical nature, but nevertheless based on syntax, not just on the linear distance between two words in a text.

The brilliant thing about Sadler's invention is that within a single dynamic data structure - a source text corpus linked with its translation - three types of knowledge are integrated: translation skill, linguistic knowledge and world knowledge. The latter refers to the semantic-proximity calculations for resolving ambiguities. No separate thesaurus or taxonomy is needed any more, and the semantic calculations rely on dependency syntax.

DLT, was it really a project of the 1980s?

In hindsight, all this appears to be in remarkable accordance with advanced insights 10-15 years later in the worldwide SMT research community. As part of a 'hybrid' approach, the use of dependency syntax has steadily gained grounds in SMT, see for example [Yamamoto 2000] and [Lin 2004]. Dependency representation is said to have the best phrasal cohesion properties [Ding 2003], and semantic dependencies are better based on syntactic dependencies than on phrasal constituents, as [Hwa 2002] points out.

[Sadler 1989] is referred to in Japanese MT texts of the early 1990s, as his analogy-based work has elements in common with the example-based work such as explored by [Sato 1990] and [Nagao 1992]. Both methods have a statistical as well as a syntactic component (Sato too used dependency syntax), but an Example-Based MT system needs a separate bank of examples and a thesaurus in addition. In 1990, I was invited to give a 2-day seminar in Kyoto about our Analogy-Based MT scheme, and in 1991 Victor Sadler himself was invited there to present a paper on the same subject.
In the western MT hemisphere, [Hutchins 1992] provides a fine and fair account of our DLT project (the documentation is flawless, apart from naming our approach 'example-based' instead of 'analogy-based'), and in their conclusions, the authors sensibly question the role of Esperanto in a BKB-based system.

It is true that the BKB model can be applied to any language pair. However, with Esperanto as one of the two languages, semantic processing may be faster, because its structural regularity enables a higher rate of transparent (parsed) syntactic dependencies, on which the semantic proximity calculations are based after all. Besides, in a multilingual system, using an IL will reduce the required number of BKBs in the same way as it would limit the number of SL-analysis and TL-synthesis modules in conventional MT. Limiting the number of different BKBs is not completely uninteresting, because each BKB requires maintenance and a lot of preparatory work. Moreover, parsing the other language in a BKB can be supported by the parse of the (less ambiguous) Esperanto version.

For Esperanto to play an important role in the SMT scene, very large volumes of present-day parallel texts in this language will be required. This is still a matter of concern, though some interesting progress has recently been made (monthly Esperanto translation of *Le Monde Diplomatique*, *http://eo.mondediplo.com*). English as a potential IL also has its pros and cons: no other language is supported by so many software resources, but which other language has such a wide-ranging diversity of geographical varieties and idioms?

Free blueprint

Packed in 1800 pages of documentation, accessible in the form of published papers and books (FORIS Publications, Dordrecht - Holland), the DLT project from the 1980s has left behind a treasure. Especially [Sadler 1989] comes down to a blueprint. There's no charge for using it, and the patents that are in place mainly serve to protect former DLT team members against any false claims by others. A call for revival of the DLT project has also been made in [Witkam 2005].

References

Bakel, J. van [1984]: 'DLT / Distributed Language Translation'.
Book review in: *Informatie*, Volume 26 #12, pp. 1010-1011.

Ding, Yuan / Daniel Gildea / Martha Palmer [2003]: 'An Algorithm for Word-Level Alignment of Parallel Dependency Trees'. Paper presented at MT Summit IX (New Orleans).

Harris, Brian [1988]: 'Are you Bitextual?'. In: *Language Technology* #7 (1988), p. 41.

Hofstadter, Douglas R. [1979]: 'Gödel, Escher, Bach: an Eternal Golden Braid'.
New York: Vintage Books, p. 601.

Hutchins, W. John / Harold L. Somers [1992]: 'An Introduction to Machine Translation'.
London: Academic Press.

Hwa, Rebecca / Philip Resnik / AmyWeinberg [2002]: 'Breaking the Resource Bottleneck for Multilingual Parsing'. Institute for Advanced Computer Studies and Department of Linguistics, University of Maryland.

IBM Europe Institute [1989]: Seminar 'Computer-Based Natural Language Translation'.
August 7-11, Garmisch-Partenkirchen.

Lin, Dekang [1995]: 'A dependency-based method for evaluating broad-coverage parsers'.
Paper presented at IJCAI-95 (Montreal), Proceedings pp. 1420–1425.

Lin, Dekang [2004]: 'A Path-based Transfer Model for Machine Translation'.
Paper presented at COLING-2004 (Geneva).

Maxwell, Dan / Klaus Schubert / Toon Witkam (eds.) [1988]: 'New Directions in Machine Translation'.
Dordrecht: Foris Publications.

Maxwell, Dan / Klaus Schubert (eds.) [1989]: 'Metataxis in Practice - Dependency Syntax for Multilingual Machine Translation'. Dordrecht: Foris Publications.

Melby, Alan K. [1986]: 'Lexical Transfer: A Missing Element in Linguistic Theories'.
Paper presented at COLING'86 (Bonn).

Melby, Alan K. [1988]: ‘Lexical Transfer: Between a Source Rock and a Hard Target’.
Paper presented at COLING’88 (Budapest).

Nagao, Makoto [1992]: ‘*Some Rationales and Methodologies for Example-based Approach’.*
In: Proceedings of the International Workshop on Fundamental Research
for the Future Generation of Natural Language Processing (FGNLP).
Sofia Ananiadou (ed.), Manchester.

Neijt, A. [1986]: ‘Esperanto as the focal point of machine translation’. In: *Multilingua,* 5-1 (1986), pp. 9-13.

Noordzij, Jaap [1987]: ‘Blackboard Translation’. In: *Systems International,* March 1987, pp. 57-60.

Papegaaij, B.C. [1986]: ‘Word Expert Semantics: an Interlingual Knowledge-Based Approach’.
Dordrecht: Foris Publications.

Papegaaij, Bart / Klaus Schubert [1988]: ‘Text Coherence in Translation’.
Dordrecht: Foris Publications.

Sadler, Victor [1989]: ‘Working with Analogical Semantics: Disambiguation Techniques in DLT’.
Dordrecht: Foris Publications.

Sato, Satoshi / Makoto Nagao [1990]: ‘Towards Memory-based Translation’.
Paper presented at COLING’90 (Helsinki).

Schubert, Klaus [1987]: ‘Metataxis: Contrastive dependency syntax for machine translation’.
Dordrecht: Foris Publications.

Tucker, Allen B. Jr. / Sergei Nirenburg [1984]: ‘Machine Translation: A Contemporary View’. In:
Annual Review of Information Science and Technology, Vol. 19, ASIS, pp. 147-149.

Tucker, Allen B. Jr. [1988]: ‘Word Expert Semantics: an Interlingual Knowledge-Based Approach’. Book review in: *Computers & Translation,* Vol. 3, pp. 83-86.

Witkam, Toon [2005]: ‘Nova vojo al Aŭtomata Tradukado’. In: Internacia Kongresa Universitato,
Vilnius, July 23-30, pp. 83-98 (*http://www.uea.org/pdf/IKU2005.pdf*).

Yamamoto, Kaoru / Yuki Matsumoto [2000]: ‘Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure’. Paper presented at COLING-2000 (Saarbrücken).

ECU (European Counting Unit) was the precursor of the euro, and was used in official transactions with the European Community Commission; its currency rate was approximately $ 1,20 towards the end of the 1980s.

Biographical details:

Toon Witkam was born in 1944. Studied aeronautical engineering at Delft University of Technology (NL). Followed a career in the software industry, dedicating the main part of it to MT. At the Dutch firm BSO in 1979 he initiated the research project DLT (Distributed Language Translation), which he led until 1989. In 1990 he was a guest researcher at ATR Interpreting Telephony Research Laboratories in Kyoto. In 1991 he advised the European Commission (DG XIII) on MT technology, and from 1992 to 1996 he was a part-time professor of informatics and cognitive ergonomics, again in Delft. He remained at BSO until 1999, being in charge of research and contacts with universities. From 2000 to 2002 he helped an internet start-up with computational linguistics, and after that he has been involved in word statistical analysis of Esperanto texts.