A new road to Automatic Translation[1]

Toon Witkam
Utrecht, Netherlands
[toon.witkam@yahoo.com]

In Hyderabad there is a big center for automatic translation. There, a battery of computers day and night reads through a mass of newspapers, magazines, books, and reports from all over the world. Mostly by internet, sometimes by scanning of paper archives. These indefatigable computers don't do that just for fun. Their human masters trained them to do this endless reading. Actually the training has been going on for decades and has become so perfect that the well-behaved machines now only support and keep up to date their translation skills by lots of reading, comparing sources in different languages treating the same topic. The ultimate dream of an ambitious translator!

Twenty years ago the grandfathers of those Hyderabad machines were still starting on the art of translation, when every day of comparative reading meant a modest improvement on the wide scale of translating experience. But the electronic sons and finally the grandsons, even faster and with a bigger memory, have acquired a level of quality higher than that of an average human translator.

So the Hyderabad center serves to support that quality level – a level that will quickly go downhill if the translation machines do not adapt to the continuous changes in the world: new terminology, new idioms, new acronyms, and – for political reasons – new geographical entities or even reviving languages. For that very reason an international team of professionals – polyglots, communication specialists, scientists – keeps watch over their fellow machine scholars. They check whether the computers are not ignoring certain sources, nor overusing others; whether they are not infected in some subtle way, whether their actual translations remain trustworthy, and things like that.

Apart from India, there are similar major centers of translation machines in the USA (San Diego), Europe (Nancy) and Korea (Pyongyang). All four centers use different machines and software, so all four operate independently of each other. Just as in the past experienced human translators did not deliver exactly identical translations of the same source text, the four centers do no do that either.
This fortunate multiplicity is skilfully exploited by internet translation services. They have at their disposal software that precisely checks whether the four translations have the same meaning. That allows them to guarantee their clients an even more faithful and high-quality translation product.

The above scenario is a projection of the state of machine translation in the year 2055. Is it science fiction? Not really. The fictitious element lies more in the organizational and entrepreneurial side of the translation centers than in the technology itself. Who is willing to make huge investments in the matter? Commercial entrepreneurs consider play robots more attractive for the market. What about governments, multinational institutions...?

[1] Originally this paper was presented in Esperanto at the 90th Universala Kongreso in Vilnius, Lithuania, 26 July 2005 (http://uea.org/pdf/IKU2005.pdf). For assistance with the English translation the author is indebted to Trevor Steele.

Technology on our plate

In the world of 2005 lightning fast machine translation (MT) is already there: through the internet, and often free of charge. The fact that those translations are not of good quality scarcely matters for most of their users. On the web we have by now got used to inaccurate, careless language use in general, so a rather poor translation is not much of a shock.

On the other hand, where texts are concerned that have to be published, whether public service announcements or handbooks used in industry, careful translation of the original, as much as possible according to a certain level or model of word stock and sentence construction, can create proper conditions for application of MT technology of a high standard. Examples of that are the Canadian translation system 'Météo' for weather forecasts, multilingual systems such as the French 'Titus' for the textile industry, the American 'Caterpillar' for export of agricultural machinery, special projects for worldwide documentation of software, etc. There are also consulting firms that adapt existing MT systems like Systran to the needs of international companies. In those cases there is high-quality translation of a 'controlled language' or 'sublanguage'.

Except for such specific applications of custom-built systems, the MT quality of publicly available services on the web is still very modest. Not just regarding the syntax and style, also regarding the choice of words. You cannot guarantee that today's translation machines faithfully reproduce the sense of the original. After a recent exploration [Hutchins 2003], an acknowledged expert on the history of MT, concluded that MT quality[2] had scarcely improved at all during the last two decades.

Please note that *the new road* described in this paper is still a course of exploration! It is about new technology which up till now did not reach the level of commercialization. What's on the market today is in fact based on research and development work done from the 1960's till the 1990's.

The lack of impressive results, after decades of research and development, is in itself quite a challenge for scientists of the new generation. New not only in age but also with regard to their discipline and methodology. The professionals who worked on MT in the years 1960 to 1990 were computer scientists, polyglots, grammarians, lexicographers, semanticists, logicians, practitioners of formal linguistics, but hardly any were professional translators. The biggest project ever in the field, EUROTRA, ended as a fiasco because of an excess of theoretical linguists [Haecken 2001].

Learning machines instead of linguists

The new generation of MT researchers formed around 1990. It grew during the last 15 years and gradually displaced the previous one. But right from the start [Brown 1988] it set off on an entirely new road. Giving up on any knowledge of linguistics at all, the new researchers[3] turned to the model of a professional *human translator*! That was almost a revolution, certainly a paradigm change. Should we now interview experienced translators, accurately observe them at work, and guess how their brains are working?

[2] Among MT systems and services tested by Hutchins were: Systran Personal, Babel Fish, Personal Translator, Lycos, Reverso, Promt, FreeTranslation, Intertran.

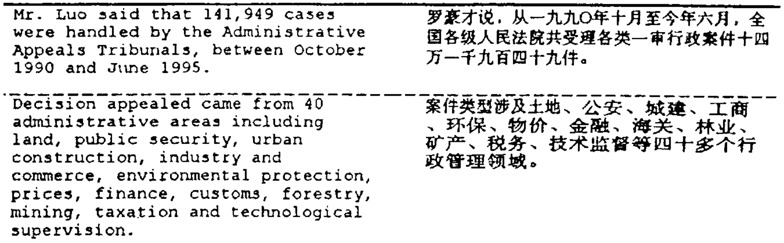
[3] Instead of linguists mainly mathematicians and software engineers.

Even more simply: should we study the results of their work, their translated texts, of course with the originals beside them? Or even more convenient, let computers do that? An excellent idea! Computers are everywhere, they get quicker and cheaper every year, they now have an enormous memory, and they don't complain about long working hours. Further, there is now the concept of "machine learning", the history of which goes back to the 1950's, about as far back as that of MT, even though the two areas did not enrich each other.

So since about 1990 the history of MT, on its very new road, is concerned to some extent with the development of intelligent[4] machines; in fact, with their training as welll as their learning, because the thing depends on having good teachers! Here are the general principles of translation learning by a computer:

- A HUGE AMOUNT of learning matter. The learning computer has to read through masses of examples of good (human) translation: at least tens of thousands, preferably millions of sentences from so-called two- or multi-language sources. Here are some:

  - Hansard (Canadian parliamentary debates),
    English and French: 2,000,000 sentences or 40,000,000 words for each language;
  - EuroParl (EU parliamentary debates, 1996-2001),
    11 languages: 740,000 sentences or 20,000,000 words for each language;
  - CNS (Chinese News Service on the web),
    English and Chinese: 25,000 sentences, 500,000 English words; e.g.:



Mr. Luo said that 141,949 cases were handled by the Administrative Appeals Tribunals, between October 1990 and June 1995.	罗豪才说，从一九九O年十月至今年六月，全国各级人民法院共受理各类一审行政案件十四万一千九百四十九件。
Decision appealed came from 40 administrative areas including land, public security, urban construction, industry and commerce, environmental protection, prices, finance, customs, forestry, mining, taxation and technological supervision.	案件类型涉及土地、公安、城建、工商、环保、物价、金融、海关、林业、矿产、税务、技术监督等四十多个行政管理领域。

- ENDLESS COUNTING. We have to explain here what is meant by the above mentioned "reading through". It means that the machine systematically works its way through every sentence-pair or sentence-multiple, registering and counting its elements (words, word sequences…) and certain relationships between them, all of this according to precise instructions given by the teacher. The counting nourishes probability calculus by which the machine subsequently (in the test phase) itself tries to translate sentences not previously seen.

- INSTRUCTION BY HUMAN TEACHERS. The researchers are also the teachers. Perhaps no longer necessary in 2055, but today still the main actors. Every MT researcher, or at least every research team, teaches the machine in their own way, experimenting with their own method or variation of a more general method. Some researchers give the computer-students very simple instructions, others teach in a much more detailed and complex way. Most often the various methods are published, and every year there are international meetings of experts. Much of that can be found on the web.

[4] In this paper the term 'intelligent machine' is used in the sense of 'learning machine'.

Above I have sketched the LEARNING PHASE. The corpus used in that phase – or the part of it used for learning – is the TRAINING CORPUS. As soon as the machine, like any learner, has to show its knowledge, the TESTING PHASE starts, and that will finally lead into the PRODUCTION PHASE.

In the test phase you present the machine with sentences from the same corpus as the training corpus. That is fair, since a corpus represents a certain type of text whose learning the exam has to test. For that reason you always divide the corpus used into two parts: the training corpus, usually the bigger part, and a smaller subset of test sentences. Of course the candidate is not to see the test sentences while learning!

Repetitive testing of intelligent translation machines, a characteristic of the new road, urges us to introduce more or less automatic judgement of the results. During the last five years several word-statistical aids[5] have been invented to measure the quality of MT products, but their use is still a matter of controversy. By the way, it is necessary to study not just the measurement of quality, but also the types of errors.

The really fast phase is – this is counter-intuitive – the learning phase, despite the huge quantity of learning material that has to be evaluated. The learning phase is indeed fully automatic, and computers are gaining speed every year.
In contrast the slowest and most expensive phase is that of preparing for instruction: the human work of researchers who keep thinking up new sets of instructions for their electronic pupils. This phase is also the most creative part of the cycle.

Those are a few of the main accents in this matter. There are some accessory phases, e.g., the boring and time-consuming previous phase of cleaning and preparing the testing corpus, of installing and making compatible the various software systems, of lining up the sentences of the corpus, etc.

Grammarless pirates as pioneers

The pioneers of the new road were mathematicians in research centers of the American firm IBM[6]. They were inspired by previous progress in the technology of automatic speech recognition – until then a research area totally unrelated to text translation. Noticing that *“The problem of language modelling for MT essentially is the same as that for speech recognition”* [Brown 1993], they elaborated a solid mathematical basis for pure SMT (Statistical Machine Translation). Their publication in a scientific journal of 1993, with 20 pages full of mathematical formulae, became the most referenced source in the field of MT. The Fundamental Equation of Automatic Translation:

$$\hat{s} = \text{argmax}_S \Pr(s) \Pr(t \mid s)$$

resumes the three challenges of SMT: estimating by computer the probability of the source-language model Pr(s), estimating by computer the probability of the translation model Pr (t | s), and thinking up an efficient and effective search method for finding that source-language word-chain which maximizes the product of those two probabilities.

[5] BLEU *(BiLingual Evaluation Understudy)*, NIST *(National Institute of Standards and Technology)*, RED *(Ranker based on Edit Distances)*, ORANGE *(Oracle Ranking for Gisting Evaluation)*.
[6] IBM Research Laboratories, Yorktown Heights, NY.

The language model is only about the linear sequence of words in a sentence, without any use of syntactical knowledge. In the translation model known as IBM Model 1 there is even a complete absence of any word sequence: instead of words from left to right in a straight line, imagine that all words belonging to the same sentence are thrown into a bag. Word order no longer exists, and the fact that by means of a trigram language model (taken over from a speech recognition system) 84% of test sentences[7] proved to be automatically reconstuctible from their word bags, is an indication of the strength of simple word statistics.

To give yourself some intuitive understanding of SMT, imagine yourself in the role of the learning machine. You (as the machine) are confronted with hundreds of thousands of sentence pairs in two languages totally unknown to you. Neither grammars nor dictionaries are at your disposal. As well, to make the exercise more amusing somebody has previously wrapped every sentence in a bag in such a way that its words[8] are totally disordered. You have in front of you hundreds of thousands of pairs of bags. You will have years for the task (in comparison, for a modern computing machine a millisecond lasts a year), so you bravely set out on this work.

You open up the two bags of the first pair and blindly you take a word from each. In the bag of Language 1 appears the word *krhŝt,* in the bag of Language 2 the word *uaaio.* You reason: "Interesting! The sentence in one bag is the translation of that in the other, so there is a chance that *uaaio* is the translation of *krhŝt."*

You check through all the pairs of bags for the presence of those two words. Obviously you count the number of times that both words occur in the same pair of bags, but you also number the "one-sided" occurrences. The relative frequency is what matters. If *krhŝt* occurs in almost every sentence it is probably a frequent function word (like *'and'* in English). If *uaaio* also occurs everywhere, that does not necessarily imply that they are translations of each other. The ideal case would be if *krhŝt* occurred in e.g., in every thousandth bag pair, always simultaneously with *uaaio* and vice versa. The result would then be: Pr (*uaaio* | *krhŝt*) = 1. More likely, *krhŝt* and *uaaio* will occur a few times without being linked if one of them has plural meanings. Then their translation probability will be calculated at e.g. 0.95, or only 0.65.

In that fashion, in the first experiment of the IBM research group [Brown 1990], the machine computed for every combination from 9,000 English and 9,000 French words[9] the probability that it is a translation-pair, and that gave a table of 81,000,000 parameters. The value of that is to give a provisional indication of word alignment, an important concept in SMT, fitly illustrated by lines between two sentences linking those words which translate each other.

However, the above sketched procedure was only the initial stage in a row of several stages: the IBM models 2 – 5 (all invented in about 1990), by iterative steps, make more precise the previously calculated probability parameters on the basis of this information: word position, fertility, and distortion.

The role of word position is obvious: translations of words in the beginning part of a source-language sentence are likely to occur in the beginning part of the target-language sentence as well, etc. A good example of word alignment perfectly in accord with word position is this:

Among the many questions raised by the expanded membership of the European Union is the question of languages.

Inter la multaj demandoj levitaj de la plivastigita membreco de la Eŭropa Unio estas la demando pri lingvoj.

In this pair of sentences the regular alignment of translated words is exceptionally lucky! Most often, translation of sentences have one or two alignment distortions, for instance because of inversion of

[7] None of the test sentences was longer than 10 words [Brown 1990].

[8] The number of words in a bag (sentence) varies between approximately 10 and 30.

[9] One had to limit the experiment to the 9,000 most frequent words in the corpus.

adjective and noun or differences in SVO (subject-verb-object) order when you translate from, say, English or Esperanto to French or vice versa. But between English and Japanese, for example, the differences in word position are much more persistent. Hence, if the word position information does not really contribute, a researcher of today uses only IBM model 1, not models 2 – 5 [Ding 2003].

A ubiquitous phenomenon is the so-called 'fertility'. When a source-language word produces a two-word rather than a one-word translation into the target language, its fertility is 2 instead of 1. A prominent example is the English function word *not*, translated into the French as *ne … pas*. The same applies to content words: there are plenty of them with fertility > 1. Look at these sentences:

Tensions | between | the | two | powers | have increased | in | recent | months.

La streĉiteco | inter | la | du | grandaj regionaj potencoj | kreskis | dum | la lastaj | monatoj.

If we take the Esperanto sentence as the source, the word "kreskis" has a fertility of 2, since it is translated as *'have increased'.* In the reverse case, the English words *'tensions'* and *'recent'* have fertility 2, while *'powers'* has fertility 3: *'grandaj regionaj potencoj'*. Most probably there are in the same text corpus other sentence pairs in which *'powers'* has different alignments, e.g. with *'grandaj potencoj'*, or simply *'potencoj'* or *'povo'*. The essence of SMT is that it catches in its probability parameters all of the variants to be found in the two-language text corpus, so in fact the products of the translator's experience and freedom – not the rules of grammar or the information in a dictionary. For that very reason SMT, the new road, differs from traditional MT.

The non-grammatical pioneers of IBM, after a corpus-based learning phase on 40,000 English-French sentence pairs[10], with an overall total of some 1,600,000 text words, achieved the following result [Brown 1990]: their intelligent machine was able to do a good translation of 48% of 73 test sentences in French. A modest success, but it was encouraging and inspiring. Their second experiment was certainly impressive [Brown 1990]: the learning machine, which had 1,778,620 sentence pairs at its disposal, calculated the translation probability of 2,437,020,096 word combinations, and by a purely statistical algorithm computed the correct alignment out of e.g. the 1.9×10^{25} theoretically possible word alignments of the following sentence pair:

What is the anticipated cost of administering and collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le côut prévu d'administration et de perception des droits?

Finally, it was another merit of the SMT pioneers of IBM at the beginning of the 1990's that they clearly agreed on the necessity of adding morphological and syntactical components to SMT in the future. The great value of their work is abiding; it was very opportune that they introduced statistical methods to the field of MT and convincingly proved their strength.

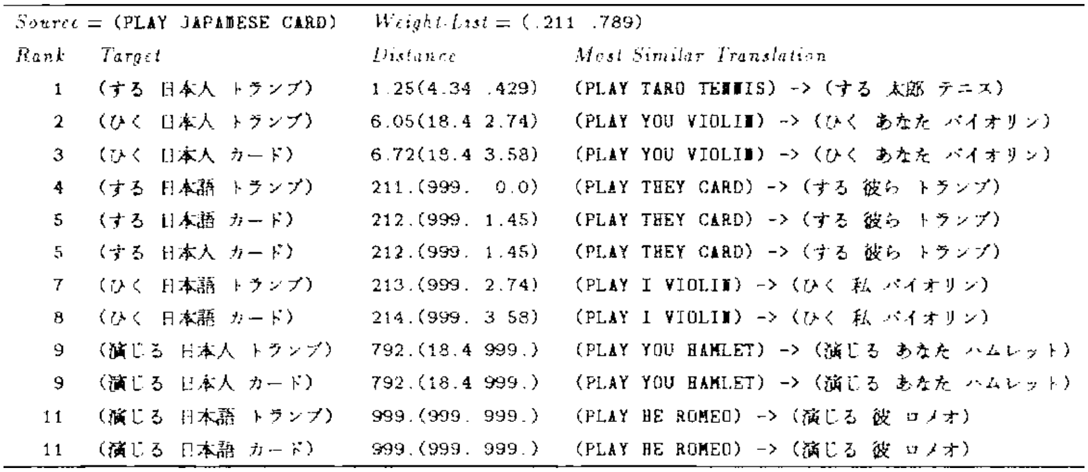
Syntax sneaks back in

Simultaneously with the developments at IBM at the start of the 1990's in the USA, but independent of them, a new paradigm was being worked out in Japan. It was close to SMT but preserved the syntax: EBMT (Example Based Machine Translation). A trait common to SMT and EBMT is orientation to texts used in the craft of translation, using bilingual corpuses or databases. Like the IBM

[10] From Hansard, the archive of parliamentary debates in Canada.

researchers the Japanese professionals were partly inspired by the work of speech recognition by computers.

The first prototypes were made by Sato [Sato 1991]. At first he experimented with a bilingual database of exemplary parts of sentences. The following table (with words re-ordered along VSO[11]) gives a quick impression of his prototype system:



Source = (PLAY JAPANESE CARD) *Weight-List* = (.211 .789)

Rank	*Target*	*Distance*	*Most Similar Translation*
1	(する 日本人 トランプ)	1.25(4.34 .429)	(PLAY TARO TENNIS) -> (する 太郎 テニス)
2	(ひく 日本人 トランプ)	6.05(18.4 2.74)	(PLAY YOU VIOLIN) -> (ひく あなた バイオリン)
3	(ひく 日本人 カード)	6.72(18.4 3.58)	(PLAY YOU VIOLIN) -> (ひく あなた バイオリン)
4	(する 日本語 トランプ)	211.(999. 0.0)	(PLAY THEY CARD) -> (する 彼ら トランプ)
5	(する 日本語 カード)	212.(999. 1.45)	(PLAY THEY CARD) -> (する 彼ら トランプ)
5	(する 日本人 カード)	212.(999. 1.45)	(PLAY THEY CARD) -> (する 彼ら トランプ)
7	(ひく 日本語 トランプ)	213.(999. 2.74)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
8	(ひく 日本語 カード)	214.(999. 3 58)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
9	(演じる 日本人 トランプ)	792.(18.4 999.)	(PLAY YOU HAMLET) -> (演じる あなた ハムレット)
9	(演じる 日本人 カード)	792.(18.4 999.)	(PLAY YOU HAMLET) -> (演じる あなた ハムレット)
11	(演じる 日本語 トランプ)	999.(999. 999.)	(PLAY HE ROMEO) -> (演じる 彼 ロメオ)
11	(演じる 日本語 カード)	999.(999. 999.)	(PLAY HE ROMEO) -> (演じる 彼 ロメオ)

Translating a new phrase (*'Japanese play card'*) from English into Japanese means calculating its semantic 'distance' from each exemplary phrase with the same verb (*'play'*). The calculation is done on the basis of a Japanese thesaurus[12], to which also English words have been added. In that way the machine finds the example closest to the phrase to be translated, and can translate it accordingly.

Makoto Nagao, the master of Japanese MT researchers, who launched the idea of EBMT already in the beginning of the 1980's, clearly explains [Nagao 1992] why this is superior to the conventional method, which depended of tedious work by linguists. As though these were lexicographers, they had to add by hand semantic indicators to every noun, exactly prescribe verb valencies, etc. That was difficult, costly, and time-consuming. On the other hand, it is simply impossible to provide enough examples to base the translation of whole sentences upon them. Nagao and [Sato 1990] guided researchers to the new road by drawing up a hybrid EBMT-framework, which makes it possible to integrate exemplary phrases into the totality of a sentence syntactical structure. Notable in that proposed framework is the use of dependency trees, instead of the constituency trees customary at that time. Also in his second, whole-sentence prototype, Sato used dependency trees.

A decade later [Yamamoto 2000] confirms the use of dependency-syntactical structures to align part sentences in SMT and implicitly in EBMT. That now helps to solve a more general problem not touched by the IBM models 1–5: the alignment of source-language word sequences to only one target-language word. The classical example of that is the English *'red herring'* and its German equivalent *'Finte'*, but there is an abundance of such non-compositional translations. To link (with only one connection line) a whole word sequence from the source language to a whole word sequence in the target language was impossible for the IBM models, whether the number of words in the two chains was equal or not. Just think of idioms and slang, just the types of sentence parts that EBMT tries to translate.

In the world of research, the hybrid translation machine, supported by SMT and syntax, is gradually gaining ground (SMT including EBMT, syntax including morphology). But there are still some zealots

[11] Verb-Subject-Object.

[12] *"Word List by Semantic Principles"*, NLRI (National Language Research Institute), Syuei Syuppan, Japan, 1964.

who are resisting the return of syntax. Koehn [2003] made a comparison of SMT results with two variations of alignment: in one method all 3-word clumps were aligned, in the other only syntactical phrases. The authors asserted that preference for syntactical structures made for poorer translations, and they challenged their colleagues who favoured syntax.

[Lin 2004], who as early as the 1990's had explored powerful parsers with the help of dependency syntax[13], took up the challenge. While Koehn et al. based their syntactical variant on constituency trees, Lin's intelligent translation machine extracted paths from source-language dependency trees of a word-aligned corpus, and translated them into fragments of target-language dependency trees. At the same time, not only the dependency relationships but also the linear sequence of words is encoded. In that way the corpus-based learning process results in a series of transfer rules with certain probabilities. After that, the translation of a new sentence develops thus: parse the sentence to acquire its dependency tree; extract from that all of the paths and find again their translations; look for a combination of transfer rules that completely handles the source-language tree and produces without conflict a target-language dependency tree; if several such combinations are found, choose the one with the highest probability.
Lin's system went through a learning phase of 116,889 sentence pairs (English-French, with 3.4 million words in total), from which 2,040,565 syntactical paths were extracted. The test phase contained 1775 sentences with a length of 5 to 15 words. Although the translation quality is still modest (BLEU score: 0.26), what is promising is the skilful transition model, whose syntax is able to handle deviations like the English-German pair *'there is' – 'es gibt'* and the English-Spanish *'swim across' – 'cruzar nadando'.*

French trees revive, American ones dry up

The new road of MT has also the characteristic that there is a growing preference for dependency syntax. That is remarkable, because for decades its big brother, constituency syntax, ruled the MT world almost alone. Here, instead of the technical details, I want to underline the almost cultural difference between the two.

Dependency syntax originated with the Frenchman Tesnière in the middle of the 20th century and gained some followers among European linguists. But when MT research developed in the United States, Chomsky's transformational-generative grammar was very influential among linguists there. That model with its abstractions and constituency syntax became a real fashion that spread among MT researchers in Europe and Japan as well. Judging by the publications of that "school", dependency syntax did not exist. That was the situation till the end of the 1980's.

We must not forget that in the world of MT and computer linguistics generally, the English language has a dominant position. The majority of research studies, systems, corpuses, parsing tools, software programs etc. relate to English. The most abundant knowledge and experience is accumulated about this language. A circumstance that contributed to it is the fact that many English-speaking MT researchers, even the (modern) linguists among them, have a very limited knowledge of "foreign" languages. The very use of the epithet ("foreign") in English-language research publications to indicate other languages reveals that.

While constituency grammar is good enough for English, whose syntactical structure is based mainly on word order (a constituent is in fact a word sequence), it is less useful for languages with a more morphology-based syntax. When dealing with a diversity of languages, dependency syntax is more suitable because it approximates to contrastive syntax [Schubert 1986].

[13] See [Lin 1995].

The tendency toward dependency trees is undeniable. According to Lopez [2002] the success of recent parsing methods [Charniak 2000; Collins 1999; Ratnaparkhi 1999] is due to ideas essentially proper to dependency syntax. Hwa [2002] confirmed that and skilfully exploited the availability of a powerful parser for English, which converts sentences into dependency trees. Such a parser does not yet exist for Chinese. By word alignment between the English and Chinese sentences of a corpus[14], Hwa (or more exactly: her intelligent machine) took word dependencies from the English side and projected them onto the Chinese side, thus creating dependency trees there. By that experiment she showed that word dependencies are more suitable for an interlanguage projection than the word-sequence constituents.

Constituency trees have still not disappeared in MT, but on the contemporary SMT road they are gradually losing their potential. Knight [2004] concedes that an alignment distortion such as in the sentence pair *'I had bought the car'* and *'Ich hatte das Auto gekauft'* cannot be handled without dependency syntax, while Koehn [2002] reported on the necessity to limit sentence length to 6 words in an experiment that aimed at constituency syntax to enrich SMT.

Finally, also as a bridge to semantics, dependency syntax performs better than constituency syntax. Hwa [2002] asserts: *"semantic dependencies form a superset based on syntactic dependencies"*, and referring to Baker [1997] she added: "work in lexical semantics research relating syntactic relationships to thematic relationships such as *agent, theme, beneficiary,* has focused primarily on syntactic dependencies rather than on phrasal constituents".

DLT results prove to be lasting

Looking backwards, to what extent does SMT relate to the former DLT[15] project? That project, which included ambitious research on MT in and from Esperanto, in fact took place before the paradigm change of about 1990, just like its competitor EUROTRA[16], which was ten times as big.

So it is all the more notable that the chief grammarian of DLT, Klaus Schubert, in the mid-1980's wisely and courageously pioneered the above mentioned tendency to dependency syntax. At a time when that method was still largely ignored in MT circles he conceived of it as the most effective method for a multi-language translation system, and published on it extensively [Schubert 1986, 1987].

As indicated above, the goal of dependency syntax in MT is to facilitate the projection or transition from elements of the source-language structure to that of the target language, i.e. contrastive syntax or *'metataxis'*, as Schubert named the process – in honour of Tesnière. Apart from that, Schubert not only described and gave the motivation for the principles of metataxis, but from 1986 to 1989 he also was active in organizing the preparation of concrete dependency syntaxes for 10 languages[17]. The results were published [Maxwell 1989].

[14] 56,000 sentence pairs of the Hong Kong News.

[15] Distributed Language Translation. DLT was a research project of the Dutch software firm BSO (1982-1990).

[16] Largest MT research project ever, in which some 300 staff members from universities throughout Europe took part, and which was financed by the European Commission (1978-1993).

[17] The languages and their syntax writers were: English (Bieke van der Korst, Dan Maxwell), Bengali (Probal Dasgupta), Danish (Ingrid Schubert), Esperanto (Klaus Schubert), Finnish (Kalevi Tarwainen), French (Luc Isaac, Dorine Tamis), German (Henning Lobin), Hungarian (Gábor Prószeky, Ilona Koutny, Balázs Wacha), Japanese (Shigeru Sato) and Polish (Marek Świdziński).

Schubert's choice for dependency syntax proved to be a solid basis, on which during 1987-1989 his colleague and head semanticist at DLT, Victor Sadler, built an avant-garde method for enabling some form of EBAT (Example Based Automatic Translation). The author himself called it *'analogical semantics'* and published his work in a book [Sadler 1989] which has been often referred to in Japanese MT papers at the start of the 1990's. While the above mentioned EBAT prototypes in Japan used a separate thesaurus to calculate semantic *'distances'* between words or phrases, Sadler's method relies only on the text corpus itself, which in its entirety functions as an example-base and a thesaurus simultaneously. That architecture put DLT on the threshold of the new SMT road. Abou this, see also Hutchins' overview [Hutchins 1992].

Semantic word distance, or *semantic proximity*, as Sadler called it, is the core of his invention. Do not confuse it with word co-ocurrence, the linear word distance (equal to the number of words-in-between plus 1) used by the technology of search tools on the internet, sometimes even by some translation systems. However, to acquire a better translation, the goal of SMT, a more subtle set of tools is necessary. To get a better idea of *semantic proximity* imagine that you urgently need a complete picture of the difference in meaning between two words as those two words are used in practice as mirrored in a large text corpus. Neither a dictionary nor a thesaurus is available, so you ask for a concordance: a list of all contexts in which word no. 1 appears[18]. If you have the memory and speed of a computer you immediately grasp that group of contexts. After that you switch into a concordance of word no. 2, and in the following microseconds you add up the differences between the two groups of contexts and deduce from that the semantic proximity of the two words: a number between 0 and 1 with two decimals. Some examples:

government	*board*	*0.89*
government	*federation*	*0.78*
government	*convention*	*0.64*
government	*communication*	*0.35*
government	*principle*	*0.27*
government	*cauliflower*	*0.11*

Here the novelty lies in the special definition of 'context': dependency-syntactic relationships[19] with neighbouring words instead of purely linear proximities, even if the latter by chance coincide with the former. The formula, by which Sadler in 1989 started teaching the learning machine to calculate *semantic proximities*, is therefore based on dependency relationships, the same as those[20] introduced by Lin more than a decade later [Lin 2004]. Enrichment of the corpus by parsing, which is demanded by Sadler's method, is feasible because parsing of phrases – detecting individual dependency relationships – is enough.

We should note that it is only the corpus that causes the two-decimal semantic numbers, measured by Sadler's dependency-based method. If by chance the corpus were a novel in which the characters were ceaselessly saying they don't trust their husbands or their government, and that their husbands and the government waste money, and so they would be happy to change their husbands and the government – in that case the semantic proximity of *'government'* and *'husband'* would possibly reach 0,90.

However, the corpus-based aspect has a great advantage over the use of a man-made thesaurus, taxonomy, or ontology[21]. Such encyclopedic structures not only necessitate continuous updating (which obviously also a corpus needs), but their adjustment implies precisely selecting the place of every new addition in a hierarchy. That is – certainly in the case of more abstract concepts – often risky and sometimes impossible. Wise supervision and equitable expansion of a text corpus as the one and only knowledge-base is not without problems, but it is at least feasible.

[18] KWIC (Key Word In Context).

[19] Subordinate relations such as 'Verb – Object', 'Noun – Adjective', 'Preposition – Noun Phrase'.

[20] Lin referred to these dependency relations as 'paths'.

[21] E.g. WordNet, EuroWordNet, and *"Word List by Semantic Principles"* (NLRI).

The semantic proximity worked out by Sadler and linked with the dependency syntax provided by Schubert, is the treasure left behind by DLT. Its value is lasting and current, because in 1989 it to a great extent anticipated the developments in the area, having created *"semantic dependencies based on syntactic dependencies"* [Hwa 2002]. Fortunately, that treasure is still accessible. Documented at great length and in detail, it has even been published[22].

In conclusion: what about a bet?

Looking forward, how much money would we be willing to bet that high-quality translation by machines will be ready in 2020, or in 2030, or …?

Certainly the memory capacity of future computers will not be a problem, nor the speed of those universal machines. Even today they are enough for almost any machine translation task. Also the provision of text corpuses (learning material for intelligent machines) is constantly growing and getting up-to-date. Actually the internet itself more and more functions as a huge multilingual corpus, and a growing number of researchers use it as such.

The new basis for SMT, in a clever hybrid arrangement with at least syntactical elements, looks healthy and promising. Compared with the rule-based rationalism of the traditional paradigm, which was excessively aimed at perfection in abstract language models, the present statistical and empirical strategy looks more suitable for gradual and constant improvement of translation machines.

Intuitively we would expect that a statistical core will make the system more flexible, like a safety net against all those unexpected cases of irregularity, including typing errors, proper names without capitals, quotes in other languages, etc. Recent progress on the new road already shows that syntactical analysis of fragments that have to be joined up later by corpus-based statistics is more successful than the eternal efforts to build a perfect parsing tool which would without fail find the one true analysis of one and every sentence. The statistical mode of operation does in a certain sense incorporate redundancy: several translations can be the result, even with negligible differences in probability. That can add strength to the translation process.

On the other hand we must not forget that the intelligent machines in test phases up till now have translated only about 50% of the presented sentences in a satisfactory manner. One way of progressing is to expand the corpuses. The bigger the text base, the more reliable the statistics. Another way is to expand and improve the various procedures (preparation of corpuses, alignment, parsing, transition, text-structural analysis).

But the most critical factor, on which depends the breakthrough to high-quality translation machines, is organizational, not technological! The researchers, scattered in their universities, naturally fond of creating ever newer variations, rarely commit themselves to common and on-going work on one sole system. For the commercial world, general and high-quality translation systems are not sufficiently attractive, and an international government like that in Brussels is afraid to risk (again) big expenses for it. Lobbying and excellent organization are necessary so that competent zealots can effectively join forces and carry out a difficult collaboration lasting many years. As a senior MT-researcher once stated [Carbonell 1992]: *"in Machine Translation, what matters is persistence"*.

[22] [Schubert 1987], [Sadler 1989] and [Maxwell 1989] can be found at www.amazon.com.

The renaissance of an (S)MT project in Esperanto, a descendant of DLT – would that not be worth a bet? Perhaps an international network or a miraculous local grouping of language-conscious computer people … competent people, for whom committed collaboration would for a time make up for the lack of a work contract in the Hyderabad center?

References

[Baker 1997] Mark C. Baker. *Thematic Roles and Syntactic Structure*. Kluwer. p. 73–137.

[Brown 1988] Peter F. Brown et al.: *A statistical approach to language translation.* Proceedings International Conference on Computational Linguistics (COLING-88). Budapest. p. 71-76.

[Brown 1990] Peter F. Brown et al.: *A statistical approach to language translation.* Computational Linguistics, junio 1990, vol. 16, number 2, p. 79-85.

[Brown 1993] Peter F. Brown et al.: *The mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, junio 1993, vol. 19, number 2, p. 263-311.

[Callison-Burch 2004] Chris Callison-Burch, Colin Bannard, Josh Schroeder: *Improved Statistical Translation Through Editing.* School of Informatics, University of Edinburgh; Linear B Ltd., Edinburgh Technology Transfer Centre.

[Carbonell 1992] Jaime G. Carbonell, Teruko Mitamura, Eric H. Nyberg, 3rd: *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, …).* Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montréal.

[Charniak 2000] Eugene Charniak: *A maximum-entropy-inspired parser.* Proceedings 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), Seattle.

[Collins 1999] Michael Collins: *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

[Ding 2003] Yuan Ding, Daniel Gildea, Martha Palmer: *An Algorithm for Word-Level Alignment of Parallel Dependency Trees.* Proceedings MT Summit IX, New Orleans.

[Hacken 2001] Pius ten Hacken: *Revolution in Computational Linguistics.* Language and Computers, december 2001, vol. 37, number 1, p. 60-72(13).

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

[Hutchins 2003] John Hutchins: *Has machine translation improved? Some historical comparisons.* Proceedings MT Summit IX, New Orleans.

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

[Knight 2004] Kevin Knight, Philipp Koehn: *What's New in Statistical Machine Translation.* Tutorial at Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Boston.

[Koehn 2002] Philipp Koehn, Kevin Knight: *ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge.* http://people.csail.mit.edu/people/koehn.

[Koehn 2003] Philipp Koehn, Franz Josef Och, Daniel Marcu: *Statistical Phrase-Based Translation.* Proceedings (Main Papers) Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton. p. 48-54.

[Lin 1995] Dekang Lin: *A dependency-based method for evaluating broad-coverage parsers.* Proceedings International Joint Conference on Artificial Intelligence (IJCAI-95), Montréal. p. 1420–1425.

[Lin 2004] Dekang Lin: *A Path-based Transfer Model for Machine Translation.* Proceedings International Conference on Computational Linguistics (COLING-2004), Geneva.

[Lopez 2002] Adam Lopez, Michael Nossal, Rebecca Hwa, Philip Resnik: *Word-level Alignment for Multilingual Resource Acquisition.* Language and Media Processing Laboratory (LAMP) Technical Report 085, Institute for Advanced Computer Studies, University of Maryland (UMIACS).

[Maxwell 1989] Dan Maxwell, Klaus Schubert (eds.): *Metataxis in Practice - Dependency syntax for multilingual machine translation.* Foris Publications.

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

[Ratnaparkhi 1999] Adwait Ratnaparkhi: *Learning to parse natural language with maximum entropy models.* Machine Learning, 34(1-3) p. 151–175.

[Sadler 1989] Victor Sadler: *Working with Analogical Semantics: Disambiguation Techniques in DLT.* Foris Publications.

[Sato 1990] Satoshi Sato, Makoto Nagao: *Towards Memory-based Translation.* Proceedings, International Conference on Computational Linguistics (COLING-90), Helsinki.

[Sato 1991] Satoshi Sato: *Example-Based Machine Translation.* Ph.D. thesis, september 1991, University of Kyoto.

[Schubert 1986] Klaus Schubert: *Syntactic Tree Structures in DLT.* BSO/Research, Utrecht.

[Schubert 1987] Klaus Schubert: *Metataxis - Contrastive dependency syntax for machine translation.* Foris Publications.

[Yamamoto 2000] Kaoru Yamamoto, Yuki Matsumoto: *Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure.* Proceedings, International Conference on Computational Linguistics (COLING-2000), Saarbrücken.