Automatic Translation: Background, Problems and Perspectives

International Colloquium on Multilingualism and Electronic Information Management, University of the Free State, 22 September 2003

Nico Weber – University of Applied Sciences, Cologne (Germany)

1 Introduction: a few terminological remarks

The term ‘Machine Translation’ – commonly abbreviated MT – is historical and polysemous. Historically it points back to pre-computer times with occasional engineering attempts at developing mechanical translating devices (Hutchins 1986; 1995). An earlier synonym, ‘mechanical translation’, is not used any longer, but ‘machine translation’ remains in common use, even if ‘computer translation’ would be more specific, and ‘algorithmic translation’ more precise, since it is algorithms running on computers that control the translation process – or its equivalent, as we will see shortly. In opposition to this, with humans in control, in MT jargon we speak of ‘human translation’. We thus have a fundamental polarity: “man translates” vs. “machine, i. e. computer translates” – with, in practice, mixed-modes between these extremes. Machine translation with human participation – also dubbed ‘interactive’ when it takes place during the translation process proper – is known as ‘human-assisted machine translation’ (HAMT); and human translation with computer support, nowadays the rule, is known as ‘machine-assisted human translation’ (MAHT) or, more commonly and briefly, ‘computer-assisted translation’ (CAT).

The term ‘machine translation’ itself is used in a broader sense and in a narrower one. In its broader sense it stands for translation performed by computer or assisted by computer. In its narrower sense it means algorithmic translation only, also referred to as ‘fully automatic translation’, in case there is no human intervention at all, or, simply, ‘automatic translation’ (AT) in the case of non-intervention into the translation process proper but of human involvement in text preparation (so-called ‘pre-editing’) and revision (‘post-editing’). The latter will be in the focus of this paper. In the following section I will outline a few characteristics of state-of-the-art AT.

2 Background: some characteristics of automatic translation

In order to characterise AT, I will only mention a few central criteria relating to its ergonomics (data processing), linguistics (language processing) and informatics (information processing). For a more detailed discussion in the light of translation theory and a systematic comparison with human translation cf. Weber (ed.) 1998.

- AT is an *“autonomous”* process, i. e. translation is executed without direct human intervention or assistance (aka ‘unassisted MT’). Indirect human assistance, on the contrary, such as selecting and preparing input documents, or updating lexical resources used by the system, is the rule. ‘Autonomous’ also means that the translation process operates according to preprogrammed rules, relying on *deductive and deterministic problem-solving (“closed or bounded”,* Sampson 1987; *“rigid processing”,* Sager 1993). This implies that problem-oriented, flexible and communicative responses *(“open-ended”, “flexible processing”)* to problems arising in the translation process are excluded. ‘Fully automatic high quality translation (FAHQT)', as it was dubbed by MT pioneer Y. Bar-Hillel, was the holy grail of MT research in its optimistic beginnings.

- AT, from users' perspective, is a *"black box"* process, which means that they have no explicit knowledge of its internal logic and workings, which they can only try to infer by studying input and output data. For various reasons system developers and vendors would rather have it this way. It has important consequences for system performance evaluation (Nübel, Seewald-Heeg eds. 1998) and lexicon editing (Kotorova, Weber 2001; Weber 2003).
- The AT approach to translation abstracts from properties and processes of the human mind and may thus be said to be concerned with *"e-language"* (external language) only – using N. Chomsky's term. We might say, alternatively, that AT processes data, not information, linguistic form and structure, not meaning and sense. This is true at least for "traditional", state-of-the-art, commercially available systems dating from the 1960s to 1980s.
- Human cognitive processes like language interpretation or understanding, and reformulation or translation at text level, in AT correspond to mainly *structure-based analysis, transfer and synthesis* at *sentence level*. The fact that there are dedicated, quasi "non-cognitive" terms for these processes is noteworthy; a terminological pendant for the whole process, however, is missing. One may wish to put "translation" in AT or MT between quotes, in order to signal that it is something qualitatively quite different from human translation.
- Automatic "translation" is *algorithmic (mechanical) data processing,* as opposed to intelligent information and knowledge processing. That is to say that AT is algorithmically controlled *language reproduction* rather than self-controlled, and self-reflexive, language production. In summary, AT is to be characterised as a mere *simulation* of translation – a "reduced version" of human translation at best, if it can be called translation at all.

2.1 Illustration: jabberwocky translation

What all this means may be illustrated by looking at an example from literature, Lewis Carroll's famous *Jabberwocky* poem in *Through the Looking Glass*[1] and by considering what it means to translate a text like this. I quote the first stanza together with two renderings into Afrikaans:[2]

JABBERWOCKY	DIE FLABBERJAK	BRABBELWOGGEL
Lewis Carroll	*Linette Retief*	*[translator unknown]*
'Twas brillig, and the slithy toves	Dis gonker en die vore garings	Dis brillig en die glyme likkedis
Did gyre and gimble in the wabe;	Fruip en gronkel in die bloof;	Drool en drindel in die weib
All mimsy were the borogoves,	Ja, grimvol was die kilderboom,	Bibberkolies is die borogis
And the mome raths outgrabe.	En die ploert wil kroof.	En die vniere rode sneib.

[1] Lewis Carroll: *Alice's Adventures in Wonderland* (1865). – *Through the Looking Glass* (1872). Harmondsworth: Puffin 1962. – I do not mean to suggest that poetry could reasonably be translated using MT. But this special – nonsensical – poem may well evidence the case in point.

[2] Cf. <http://www76.pair.com/keithlim/jabberwocky/> (08-2003) with dozens of other translations, parodies and variants of the poem. – "Die Flabberjak" by L. Retief appeared in "Die Burger", Cape Town, 25 August 1992.

Translation, like any other linguistic activity, crucially involves *making sense*. The word 'making' is to be stressed. Linguistic expressions – from morphemes or words to texts – do not "have" meaning nor do they "carry" meaning in the sense it would at any time subsist independently of individuals communicating by language.[3] Meaning is a process, not a product, the process of making sense of linguistic expressions, i. e specific symbols.[4] And there are active "processors" crucially involved in this process who are able to interpret the expressions or symbols concerned. Artificial systems until now can only simulate interpreting symbols and making sense of them. Since human cognitive systems are black boxes just as automatic translation systems are, one cannot directly observe their making sense, but one can conclude that they do from their reactions and behaviour. A relatively transparent, observable mode of making sense is translation. A translated target text is the visible result of a translator's having made sense of the source text.

In the light of what has been said the *Jabberwocky - Flabberjak - Brabbelwoggel* texts above turn out to be examples of non-translations. The source text being deliberately nonsensical, it is not possible to make sense of it, at least not in the usual way relevant in communication. The target texts are not translations but some kind of imitations or, if you like, results of simulated translation, the nature of which we need not discuss in detail here. However, not everything is nonsensical in the source text. There are features and forms taken from English, such as punctuation, bound and unbound morphemes, i. e. affixes and function words – auxiliaries *(was* and *did),* articles *(the),* conjunctions *(and),* prepositions *(in)* – and, most significantly, word order and phrase structure. Looking at the renderings of the source text in Afrikaans – as well as in other languages – we can see that these structures are reproduced – with more or less variation.

And this is a close approximation of how AT works. It takes its cues mainly from text-encoded source-language morpho-syntactic elements and structures, reproducing or transforming them into target-language equivalents declared to be corresponding in its internal set of rules (so-called transfer rules). Kay et al. 1994, 63 characterise the process of AT as "mapping source language representations to target language representations." Note that the "translation mapping" is done not between source and target language but between source and target language *representations*. Lexical equivalents would be encoded in the system lexicon(s) and substituted in the same way. Nothing like "making sense" of them is involved. If we recorded equivalents for expressions like *brillig, slithy, tove, gyre, gimble,wabe* in a bilingual AT system for English to any other language, the example sentences would be "translated" like any other.

2.2 Three examples of automatic translation

We will now have a look at three small AT samples submitted to the Internet-based ALTAVISTA "Babel Fish Translation" service powered by SYSTRAN.[5] The texts, also from the WWW, are in

[3] This is the "conduit metaphor" criticised by: Reddy, Michael J. 1979: *"The Conduit Metaphor – A Case Frame Conflict in Our Language about Language"*. In: A. Ortoni (ed.): Metaphor and Thought (284-324). Cambridge: U. P.

[4] It is often helpful, in my opinion, to interpret the term 'meaning' (and several others) as *nomen actionis* (action or process noun) instead of *nomen acti* (result or static noun).

[5] <http://babelfish.altavista.com> (08-2003).

three different source languages (French, German and Russian); the target language was English in each case. They deal with the same subject (automatic translation) and were googled on the WWW using the search words *traduction automatique, automatische Übersetzung, avtomatičeskij perevod*.

French-language source text [1] is extracted from YAHOO Encyclopedia *(YAHOO Encyclopédie)*[6] – headword: *la traduction automatique* (automatic translation).

L'approche linguistique de la traduction automatique

L'un des domaines de la linguistique appliquée est celui de la traduction, en particulier la traduction automatique. La multiplication des ordinateurs a laissé espérer la possibilité de remplacer le traducteur humain par une machine, ce qui impliquait des descriptions formelles de la syntaxe et de la sémantique des langues concernées. De ce point de vue, les travaux du linguiste américain Noam Chomsky, qui partait de l'hypothèse qu'il y a des structures communes à toutes les langues, ont paru un temps prometteurs, mais on s'est aperçu qu'on ne pouvait pas transposer de façon automatique une langue dans une autre, et qu'il était nécessaire de passer par une sorte de langue intermédiaire, de caractère universel. Ces travaux ont ouvert la voie à des recherches concernant la linguistique mathématique et les universaux du langage, mais les résultats sont pour l'instant limités.

Linguistic approach of machine translation

One of the fields of linguistics applied is that of the translation, in particular machine translation. The multiplication of the computers let hope for the possibility of replacing the human translator by a machine, which implied formal descriptions of syntax and semantics of the languages concerned. Of this point of view, work of linguist American Noam Chomsky, which left the assumption that there are structures common to all the languages, appeared a promising time, but one realized that one could not transpose in an automatic way a language in another, and that it was necessary to pass by a kind of intermediate language, of universal nature. This work opened the way with research concerning mathematical linguistics and the universals of the language, but the results for the moment are limited.

The above translation is surprisingly good by AT standards; at least it is perfectly understandable. Problems in the target text that would have to be amended are underlined. The reason for this relatively high quality is structural similarity: the target rendering in English of the French source is quite literal, which is most obvious by inadequacies like *linguistics applied < linguistique appliquée* or *work of linguist American < les travaux du linguiste Américain*. That it is not a simple word by word translation is demonstrated by the diverging use of articles and number, and by a case like *a promising time,* an inverted rendering of *un temps prometteurs* – the correct rendering would have been *promising for a time,* to which a literal rendering **a time promising* would have happened to be semantically closer.

[6] <http://fr.encyclopedia.yahoo.com> (08-2003).

German-language source text [2] is extracted from a short survey article on *Machine Translation (Maschinelle Übersetzung)* by LINGUATEC Sprachtechnologien GmbH.[7]

Simple Übersetzungsprogramme orientieren sich nur an der Oberflächenstruktur und übertragen deshalb einfach ein Wort nach dem anderen. Dabei kann natürlich nur Kauderwelsch (…) herauskommen. Brauchbare Ergebnisse liefert nur Software, die Satz für Satz zuerst die Oberflächenstruktur eines Satzes in der Ausgangssprache analysiert, sie kategorisiert, um die darunter liegenden Tiefenstrukturen zu erkennen. Diese überträgt sie dann in entsprechende Strukturen in der Zielsprache. Auf die satzweise Analyse beschränkt man sich aus praktischen Erwägungen. Wünschenswert, aber derzeit technisch nur in wenigen Übersetzungsprogrammen realisiert, ist eine Analyse von größeren Einheiten, idealerweise satzübergreifend.

Simple translation programs orient themselves only at the surface texture and transfer therefore simply a word after the other one. Naturally only Kauderwelsch (?) can come out. Only software, which analyzes sentence for sentence first the surface texture of a sentence in the source language, supplies useful results it categorized, in order to recognize the depth structures which are under it. This transfers it then into appropriate structures in the target language. To those analysis is limited sentence by sentence one from practical considerations. Desirably, but at present technically only in few translation programs realized, is an analysis of larger units, ideal-proves satzuebergreifend.

Here the target text is distinctly less comprehensible. Of course, modules responsible for translating different language pairs in an AT system like SYSTRAN are autonomous from a linguistic point of view and results may thus qualitatively diverge. It is obvious that sentence structures in German, as exemplified by source text [2], differ from the syntax of English to a greater extent than does the syntax of French. If basic word order is SOV, as in the first sentence *(Simple Übersetzungsprogramme ...)*, it is mirrored in the target sentence. If it is OVS it is rearranged into SOV, as in the third sentence *(Brauchbare Ergebnisse ...)* – a case where the process is enormously complicated and leads to a labyrinthine rendering, to say the least, due to a complex overall structure with second-level nestings. The fourth sentence, which is SOV, is reproduced word by word, the result being at least stylistically objectionable. Unknown words in the source text (here: *Kauderwelsch* and *satzübergreifend)* may impede the syntactic analysis of the sentence in which they occur.

[7] In the Internet magazine *aboutIT* of 23. November 2002, published by Stefan Bachert GmbH – <http://www.aboutit.de> (08-2003).

Russian-language source text [3] is extracted from a brief *History of Machine Translation (История машинного перевода)* by A. E. Vladimirovna.[8]

К началу 50-х годов целый ряд исследовательских групп в США и в Европе работали в области МП. В эти исследования были вложены значительные средства, однако результаты очень скоро разочаровали инвесторов. Одной из главных причин невысокого качества МП в те годы были ограниченные возможности аппаратных средств: малый объем памяти при медленном доступе к содержащейся в ней информации, невозможность полноценного использования языков программирования высокого уровня. Другой причиной было отсутствие теоретической базы, необходимой для решения лингвистических проблем, в результате чего первые системы МП сводились к пословному (*word-to-word*) переводу текстов без какой-либо синтаксической (а тем более смысловой) целостности.

A whole series of research groups in THE USA is annual at the beginning of the 50th and in Europe they worked in the region MP. In these studies significant means were inserted; however, results very soon disappointed investors. One of the main reasons for low quality MP in those years were the limited possibilities of the hardware: the small storage capacity with the slow access to the being contained in it information, the impossibility of the valuable use of languages of programming high level. The absence of the theoretical base, necessary for the solution of linguistic problems, was another reason, as a result of which the first systems MP were reduced to word-by-word (*word- that -word*) transfer it was text without any syntactic (*but that more* semantic) integrity.

Again we can observe a distinct tendency in the target text to reflect source text syntactic structures *(to the being contained in it information < к содержащейся в ней информации)*. Moreover, there are a number of lexical problems (dotted underlines), e. g. *in the region MP* instead of *in the domain of MT < в области МП;* or lexico-syntactic ones, e. g. *languages of programming high level* instead of *high-level programming languages < языков программирования высокого уровня;* or *but that more* instead of *a fortiori* or *let alone < а тем более*. We could say that the target text is reasonably understandable, but far from acceptable by high quality translation standards.

The examples presented in this section convey an impression of the quality level that can be expected from AT. Of course this is a more or less random impression, since only one AT system and three translation pairs were chosen. However, SYSTRAN is one of the oldest and most used AT systems, and the four selected languages are among the most relevant for AT (together with Japanese and Spanish).

3 Problems and Perspectives

3.1 Problems: why AT is difficult

This problem has been discussed from the beginning of MT research and development; I here refer to Sampson 1987 and Kay et al. 1994, 11 ff. as two of the most interesting contributions. One obvious reason why AT is difficult is that translation is difficult. Another reason, which we have touched on above, is that AT de facto is not even translation.

Sampson 1987, 92 argued that there is nothing “identifiable as ‘a 100% faithful [or accurate] translation’, which clever humans might manage to produce.” Also, in many cases, there is no unique way for a sentence, let alone a text, to be translated. In the same manner as the meaning of an expression does not “exist” independently of users and context, its translation is not, as it

[8] <http://www.langust.ru/etc/history.shtml> (08-2003).

were, “pre-established”, and cannot be “deduced” from the source, but it has to be “made up” in context. That means that deterministic, deductive procedures are inadequate for solving translation problems.

Kay et al 1994 underlined the roles of context and convention. The notion of *context* covers two aspects: the linguistic or textual context and the extralinguistic or situational context.[9] This distinction and its implications have been elaborated by (London School or British) contextualism under J. R. Firth and his followers. The point, briefly stated, is that linguistic expressions do not “have” (much) isolated meaning,[10] but that their meaning “develops” in the context of other expressions and a specific situation. This situation is either one of “encoding”, i. e. of producing or uttering meaning, or one of “decoding”, i. e. of reproducing – in the sense of: “recreating” – meaning, or understanding. Both situations – uttering and understanding – imply generating meaning in context. The implication for translation once again is that a translation cannot be “deduced” from the source sentence or text, as conventional algorithms would try to do, but must be “induced” and “originated” using linguistic and extralinguistic contextual information.

Another, closely related, factor is the role of *convention* which entails that much of what is to be understood in a text is not explicitly stated but has to be implied. Inversely, “much of what is present in the utterance of the text is there, not because it is essential to the message, but because it is required by the language or the culture.” (Kay et al. 1994, 22). You need to be aware of these things in order to use languages competently and produce good translations, which state-of-the-art MT systems manifestly are in no way. Sampson 1987, 92 concluded that “what MT needs to do is to look for tricks that work more often than not and that, even when a mechanical translation is identical to what a human translator might have written, is produced by a quite different method.” Kay et al. 1994, 85 point to this difference by concluding that translation is not “a function from a source to a target text. (…) translation is essentially a process in which information is lost and gained: translation is not meaning preserving.”

3.2 Perspectives: how to get along

Simply put, the question is, what to do if AT is not as good as it should be? Of the alternative

- either to make do with less
- or to seek to make it better

both choices have been opted for. The *“make do with less”* strategy means that less than perfect results are traded for advantages like low or even no cost, all-time availability and speed. Free Internet AT services are a well-known example. We have seen a few examples from ALTAVISTA/ SYSTRAN in section 2.2 above. This has been dubbed ‘indicative’, ‘information-only’ or

[9] There are further sub-categorisations which are important but cannot be discussed here: 1. The linguistic context includes word environment (narrower sense) as well as broader textual context (broader sense). In text corpus analysis these are captured by the concepts of syntactic colligation and lexical collocation vs. lexical association. 2. The situational context encompasses everything from the immediate participants and setting of an utterance (narrower sense) to the global cultural context of a language (broadest sense).

[10] This formulation is somewhat oversimplified. What is meant is that linguistic expressions taken for themselves are vague and, often, ambiguous (“indeterminacy”). They become as precise and disambiguated as necessary in “situated” language, i. e. language used in context.

‘gist(ed)’ translation.[11] Its usability should not be assessed from the perspective of someone familiar with both source and target languages but, rather, of someone who does not know the source language (possibly including its script) at all: is the target text able to convey a reasonable idea of what the source text is about?

The second option is to use AT in the context of intellectual pre- and post-editing. It has been practised for some decades in a number of – mostly larger – organisms and companies (cf., e. g., Kay et al. 1994, 40 f.; Hutchins 1999) although, generally speaking, the role of AT is not as important by far in industrial practice, including the so-called language industries, as that of human-assisted MT (translation memories). It is common opinion AT works best with certain types of text. The best-known and, in a certain sense, extreme, example for special text AT is the Canadian TAUM-METEO English⇔French weather bulletin translation system. Weather bulletins are linguistically characterised by a very reduced vocabulary used in a small number of stereotypical phrases.

The *“seek to make it better”* strategy, of course, is a driving force of AT research and development. I cannot go into many details here. I only want to name four general domains on which further development of AT has concentrated during the last 10-15 years.

- Integration of formal linguistic theory, i. e. computationally oriented, declarative syntax frameworks – particularly unification-based formalisms (Kay 1994, 56 ff.).
- Non-linguistic approaches (1), including (a) statistically based AT; (b) “example-based”[12] MT (Kay 1994, 64 ff.); (c) connectionist approaches based on radically different data processing concepts and techniques.
- Non-linguistic approaches (2): AI- or knowledge-based MT which seeks to model extralinguistic or “world knowledge” as well as logical and so-called “common-sense reasoning”, in order to make the AT process more “realistic” (Kay 1994, 72 ff.).
- Development of existing AT systems and resources. The most important linguistic resources are sets of rules, the grammar, which is used mainly in analysis and synthesis, and sets of interlingual lexical equivalence declarations, the lexicon(s), chiefly used in transfer. The lexicon component, as a rule, is the only interface – apart from text input and output – accessible to users. A principled and systematic study of the lexicon component of MT systems, and their maintenance and enlargement in order to enhance translation quality, has been dubbed *MT lexicography.* It can be viewed as a sub-discipline of computational lexicography.

3.2.1 SYSTRAN

According to Wilks 1992, for whom “SYSTRAN (..) remains the existence proof of machine translation” (166),“the power of SYSTRAN lies in its well-established and relentless system of lexicon modification and augmentation in the face of bad translation results.” (169). SYSTRAN is one of

[11] Cf.: *Automated Real-time Translation (ART) and the Power and Purpose of „Gisting“ in the Internet Era.* A White Paper from Transparent Language, Inc. February 2001: <www.TransparentLanguage.com> (06-2001). – Alternative expressions for ‘gisting’ are: ‘information scanning’ or ‘assimilation’.

[12] A better term for this approach, in my opinion, would be “pattern-based MT”.

the oldest operational AT systems.[13] It was developed for Russian-English at Bonn University (Germany) from 1964 by Peter Toma, who had already been working on Russian-English MT at Georgetown University (USA) for more than a decade. The Russian-English version was first used by the US Air Force in 1970. In 1974 Toma started development of an English-French version, demonstrated to the European Commission in Luxembourg in 1975. The Commission contracted the right of development for this and other language pairs in 1976 – from French-English in 1977 to 16 other language pairs (by unidirectional count) until 1997. Note that the Internet version as well as commercially available PC versions are not identical with those developed by the European Commission. The latter, in contrast, are not available to the public.

The SYSTRAN lexicon system is described in detail by Schäfer 2002, 41 ff. with further references). Summarising a larger-scale empirical evaluation of the Commission's SYSTRAN for French ⇔ German based on two special-language-of-economy corpora with 10 000 running words each, Schäfer 2002, 298 concludes that although the SYSTRAN lexicons have been upgraded and enlarged for many years to a size and richness probably unattained by any commercially available system, much remains to be done in order to further enhance translation quality for special-language texts like those considered in his study. He declares himself sceptical about the possibility of further progress, "given the nearly infinite diversity of linguistic structures [and] the current state of MT research" (my translation).

3.2.2 MT lexicography

MT lexicography is indeed a challenging R&D topic. As two empirical studies (Kotorova, Weber 2001; Weber 2003) involving several commercial AT systems have shown, a majority of translations that on review are qualified as incorrect or inaccurate[14] have to be imputed to problems of interlingual lexical matching, and most others to syntax – some also to both. Many of these are target language collocational errors or interlingual head (category) switching mismatches (Eberle 2001). In order to handle these problems we would again have to distinguish the two context types mentioned above (3.1).

Linguistic context, on the one hand, is evidenced by *collocational relations* and *selectional preferences*.[15] From an engineering perspective on interlingual mapping Kay et al 1994, 73 maintained that: "The problem stems from the fact that lexical selection restrictions are tied to lexical items and not to concepts. It is therefore necessary to apply them twice, once to source structures and once to target structures." The *context of situation,* on the other hand, is represented by discourse forms and functions. We may think of the famous "language games (Sprachspiele)", devised by the philosopher L. Wittgenstein, and defined in the *Philosophical Investigations* as: "the totality of language and all activities with which it is 'interwoven'" (my translation).[16] Transla-

[13] Historical details based on Hutchins 1986 and Schäfer 2002

[14] This deliberately vague formulation is not meant to hide the problems behind it as, e. g., the questions of standards of evaluation. It is simply not possible to go into this here.

[15] I prefer the term 'selectional preferences' to 'selectional restrictions' because it is less absolute and more to the point – cf. Wilks, Yorik 1975: *"Preference Semantics"*. In.: E. L. Keenan: Formal Semantics of natural Language (329-348). Cambridge: U. P.

[16] „Ich werde auch das Ganze: der Sprache und der Tätigkeiten, mit denen sie verwoben ist, das 'Sprachspiel' nennen.“ Wittgenstein, Ludwig 1958: *Philosophische Untersuchungen*. Frankfurt/Main: Suhrkamp (^{3}1982, §7).

tion, in this perspective, is about devising analogous "language games" in a target language for those represented in the source text. As Kay et al 1994, 94 put it: "The key point is that capturing analogous discourse-function seems to be far more important than mimicking the source language syntax". A central problem is how to represent contextual information of all types in AT system lexicons of in order to make it operational in translation.

To quote an example (from Weber 2003) of seemingly straightforward lexical mapping: in order to translate the noun *guide* in English into German, you have to decide

(1) whether it designates a person *(Führer)*

(2) or a text, giving information about something, or assistance in assessing something – in this case, you have to decide

  (2-1) whether it designates a book

    (2-1-1) either pertaining to tourism *(Reiseführer)*

    (2-1-2) or to some other, e. g. technical, domain *(Handbuch)*

  (2-2) or some other text form which, then, you do not specify *(Anleitung, Ratgeber, Richtschnur,…)*

(3) or a sign indicating direction or path *(Wegweiser, Wegmarkierung)*.

As a problem of interlingual lexical equivalence (mapping) this is well-known of course. What has not been discussed in so much detail is the quality and status of the equivalents quoted in the example above. These actually are not so much representations of lexical elements than of conceptual elements. Lexical items do not "stand for themselves" but for sets of expressions which, in some cases, may also be one-element sets.

1. In identical contexts they can in many cases be replaced by synonyms (e. g., *Führer* by paronyms like *Fremdenführer, Reiseführer)*.
2. In nonidentical contexts, in contrast, they cannot be replaced by synonyms (e. g. by equivalents of *conductor, leader* or *teacher* – all three potential equivalents of *Führer)*.
3. With polysemous expressions like *Reiseführer* (a case of metonymy: it may designate either a person or a textbook) translation implies a specification of which meaning "branch(es)" are involved. Translation, as is well known, implies disambiguation.[17]

It turns out that, at least for AT purposes, lexical units frequently are not the best way of representing matching conceptual units. An alternative may be sets of lexical units, but these are often not closed. Elements of different formal representations can be found in various AT dictionary interfaces, but there is no universally applicable solution.

3.2.3 Universal Networking Language

A major international effort to tackle these problems is the *Universal Networking Language (UNL)* project, co-ordinated by the *United Nations University/Institute of Advanced Studies (UNU/IAS)* in Tokyo starting in 1996 and, since 2001, the *UNDL Foundation* in Geneva.[18] UNL is an artificial language, more precisely, a formal meaning representation language devised to

[17] In the sense of becoming aware of "branches of meaning", including the recognition of that the equivalent may be identically polysemous or ambiguous.

[18] <www.unl.ias.unu.edu> – <www.undl.org> (08-2003).

function as an *interlingua* for use in AT. In MT R&D the interlingua concept, rooted in rational philosophical tradition, holds the promise to overcome the deficiencies of AT essentially based on structural transfer. The idea is to bring AT closer to how human translation works by founding transfer on meaning instead of lexico-syntactic structure. In the first place this approach is confronted with problems of representation. As linguists are well aware there is no universally acknowledged method of representing meaning. Natural languages are well-tried meaning representation systems; the problem is, though, that they differ in coverage and in many details. An interlingua, thus, should be able to represent everything that either of the languages between which it is devised to "mediate" can express. The problem has not been solved to-date. A compromises between the structural transfer approach and the semantic interlingua approach is dubbed a *pivot language*.

UNL is a pivot language based on what Kay et al. 1994, 96 called a "componential strategy", since it "attempts a decomposition into primitive conceptual components".[19] It consists of "universal words (UWs)" – its vocabulary – that can be linked by relations and modified by attributes – its syntax. UWs represent interlingual concepts ("conceptual labels"); they are defined in the UNL knowledge base – the semantics of UNL. The knowledge base is a hierarchically organised, inheritance-based network of representations of conceptual units and their relations and attributes. UW representation is based on English; forms that are ambiguous in English are disambiguated and forms that do not exist in English can be declaratively added to the UWs set.

The engineering concept is to develop UNL-based Internet browser add-ins dubbed 'enconverters' and 'deconverters' which translate Internet content from a natural language into UNL and, inversely, from UNL into another natural language. The translation would, in the "classical" MT approach, proceed sentence by sentence, with a combination of UWs representing the meaning of each sentence. The interface between natural language lexical elements and UWs in UNL is encoded in special dictionaries. This is a very interesting project and it remains to be seen what will finally result. If it works it will be a real breakthrough in MT R&D.

3.3 Summary

After introducing a few central terms and concepts related to machine translation (MT) (1) this paper focussed on (fully) automatic translation (AT). I gave some background information on AT by briefly discussing its main characteristics (2) followed by illustrative examples (2.1 and 2.2). The point was to make clear that AT is a kind of reduced version, or a mere simulation, of translation. Knowing that and its reasons one is able to assess AT results, their deficiencies and limits in an informed and realistic way (3.1): AT is difficult because translation is difficult and even more so because it is not even "real" translation. I then revised different strategies to be adopted as a consequence (3.2) – basically, the "make do with less" strategy on the one hand and the "seek to make it better" strategy on the other. The first strategy consists in either operating with AT results as they are – alternatively, in improving them by human intervention – or in trying to select or manipulate input data in order to make them as suitable as possible for AT processing. The second strategy is a driving force of AT research and development. As an illustration

[19] The following brief exposition is based on Uchida, Zhu 2001 and Zhu, Uchida 2002. – Both papers are downloadable at <http://www.unl.ias.unu.edu/publications>.

I presented three paradigmatic examples, namely: SYSTRAN as an example for a long-time "pragmatic" approach consisting in continuous "case-based" improvement of an AT system (3.2.1); MT lexicography as an array of efforts to improve and extend the lexical (and grammatical) information base of existing AT systems based on newer linguistic theories and insights (3.2.2); finally, the "Universal Networking Language" project as a promising new large-scale development under the interlingua or pivot language paradigm (3.2.3).

4 References

Eberle, Kurt 2001: *"FUDR-based MT, head Switching and the Lexicon"*. Proceedings of the MT Summit VIII, Santiago de Compostela, 18-22 September 2001.

Hutchins, W. John 1986: *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood (New York: Halsted).

— 1995: *"Machine Translation: A Brief History"*. In: E. F. K. Koerner, R. E. Asher (eds.): Concise History of the Language Sciences: From the Sumerians to the Cognitivists (431-445). Oxford: Pergamon.[20]

— 1999: *"The Development and Use of Machine Translation Systems and Computer-based Translation Tools"*. In: Ch. Zhaoxiong (ed.): Proceedings of the International Conference on Machine Translation and Computer Language Information Processing, Beijing, 26-28 June 1999 (1-16). Beijing: Research Centre of Computer & Language Engineering, Chinese Academy of Sciences.[20]

Kay, Martin, Jean Mark Gawron, Peter Norvig 1994: *Verbmobil: A Translation System for Face-to-Face Dialog*. CSLI (Lecture Notes No. 33).

Kotorova, Elizaveta, Nico Weber 2001: *"Interlingual Lexical Equivalence in Machine Translation"*. In: P. Kocsány, A. Molnár (eds.): Wort und (Kon)text (15-47). Frankfurt/Main: Lang.

Nübel, Rita, Uta Seewald-Heeg (eds.) 1998: *Evaluation of the Linguistic Performance of Machine Translation Systems*. St. Augustin: Gardez!.

Sager, Juan C. 1993: *Language Engineering and Translation: Consequences of Automation*. Amsterdam, Philadelphia: Benjamins.

Sampson, Geoffrey 1987: *"MT: A Nonconformist's View of the State of the Art"*. In: M. King (ed.): Machine Translation Today: The State of the Art (91-108). Edinburgh: U. P.

Schäfer, Falko 2002: *Die maschinelle Übersetzung von Wirtschaftsfachtexten: Eine Evaluierung anhand des MÜ-Systems der EU-Kommission, SYSTRAN, im Sprachenpaar Französisch-Deutsch* [Machine translation of special texts in the domain of economy: an evaluation of the EU Commission's MT system, SYSTRAN, in the language pair French-German]. Frankfurt am Main [etc.]: Lang.

Uchida, Hiroshi, Meiying Zhu 2002: *"The Universal Networking Language Beyond Machine Translation"*. International Symposium on Language in Cyberspace, Seoul, 26-27 September 2001.[21]

Weber, Nico (ed.) 1988: *Machine Translation: Theory, Applications, and Evaluation*. An assessment of the state-of-the-art. St. Augustin: Gardez!

— 2003: *"MÜ-Lexikographie"* [MT lexicography]. In: U. Seewald-Heeg (ed.): Sprachtechnologie für die multilinguale Kommunikation: Textproduktion, Recherche, Übersetzung, Lokalisierung (145-183). St. Augustin: Gardez!.

Wilks, Yorick 1992: *"SYSTRAN: it obviously works but how much can it be improved?"*. In: J. Newton (ed.): Computers in Translation (166-188). London, New York: Routledge.

Zhu, Meiying, Hiroshi Uchida 2002: *"Universal Word and UNL Knowledge Base"*. ICUKL-2002, Goa, 25-29 November 2002.[22]

[20] Also on the Internet at: <http://ourworld.compuserve.com/homepages/WJHutchins>.

[21] <http://www.unl.ias.unu.edu/publications/UNL-beyond MT.html> (09-2003).

[22] <http://www.unl.ias.unu.edu/publications/UW and UNLKB.htm> (09-2003).