MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society
Issue No. 12
December 2001

The *Machine Translation Review* incorporates the Newsletter of the Natural Language Translation Specialist Group of the British Computer Society and appears annually in December.

The Review welcomes contributions, articles, book reviews, advertisements, and all items of information relating to the processing and translation of natural language. Contributions and correspondence should be addressed to:

Derek Lewis
The Editor
Machine Translation Review
School of Modern Languages
Queen's Building
University of Exeter
Exeter
EX4 4QH
United Kingdom

Tel: +44 (0)1392 264325
E-mail: D.R.Lewis@exeter.ac.uk

ISSN 1358–8346

Please note: From April 2000 this Review will be published electronically and will be available on our web site at the British Computer Society (see page 5). The format will be in HTML, in the same way as some back copies have already been stored electronically, so it will be easy for readers to print copies if they wish. Each section will be separate so readers may print selected parts only.

Some copies will be printed for the Copyright libraries and for purchase at a modest price plus postage and packing for those without electronic access. Members of the Natural Language Translation Specialist Group of the British Computer Society will be advised of each issue.

Contents

Group News and Information 4
Letter from the Chairman 4
The Committee 5
BCS Library 5
Website 5

Latin to English Machine Translation - A Direct Approach
Paul R. Bowden 6

Learning Machine Translation Strategies Using Commercial Systems: Discovering Word Reordering rules
Mikel L. Forcada 13

An Example Based MT System in News Items Domain from English to Indian Languages
Dr. Sivaji Bandyopadhyay 19

Towards a Fully-Automatic High-Quality Machine Translation System for Unrestricted Text
Dr Mirosław Gajer 24

PC-Based Machine Translation: an Illustration of Capabilities in Response to Submitted Test Sentences
Derek Lewis 36

Semi-Automatic Construction of Multilingual Lexicons
Lynne Cahill 58

Seamless Translation - Delivering for the User
Phil Scanlan 67

Book Review 75
Conferences and Workshops 77
Membership 80

Group News and Information

Letter from the Chairman

The British Computer Society
Charity No. 292786

THE NATURAL LANGUAGE TRANSLATION SPECIALIST GROUP

Please reply to:
72 Brattle Wood
Sevenoaks
Kent, TN13 1QU
Tel: 01732 455446
E-mail: wiggjd@bcs.org.uk

26 January 2002

'As you may already know, this Review is being published on our website from which printed copies can be obtained easily. However, we have produced some printed copies for Libraries and for sale at a modest price of £2.00 each for those people who would like printed copies (Please see our website for details of cost of postage etc.).

When we set up an e-mail list (bcs-mt@jisc.ac.uk) at http://www.jisc.ac.uk earlier this year it was thought that this would replace our membership records at BCS Headquarters in Swindon, but it has since transpired that they are able to provide us with lists of e-mail addresses as well from their records so we now intend to continue to maintain these records. However, the e-mail list , bcs-mt@jisc.ac.uk remains available for members to communicate queries and problems with other members. To reach a wider international audience we recommend the use of mt-list@eamt.org (at http://www.eamt.org).

Once again we must ask you to consider contributing to this Review. We would still welcome more articles, papers and reports on the subject of machine translation and related subjects such as computer assisted language teaching, computer based dictionaries and aspects of multi-linguality in computing etc. We would welcome papers from staff and students in linguistics and related disciplines, and from translators and any other users of MT software.

May I remind members yet again, that they do not need to live near London to assist the Committee. We do not have sufficient funds to pay travel expenses for all Committee members to attend meetings, but we still welcome Correspondent members. Correspondent committee members are otherwise treated as full members of the committee and kept advised of all committee business. Anyone interested in helping should contact me or any other Committee member.

Our committee still requires a treasurer, in our case the role is more of an auditor since all our transactions are processed by the BCS. This post does, of course, require some knowledge of accounting, but not much I'm glad to say, and, as mentioned above this does not need to be for someone in the London area. Anybody interested to know more, please contact me.

All opinions expressed in this Review are those of the respective writers and are not necessarily shared by the BCS or the Group.'

The Committee

The telephone numbers and e-mail addresses of the Committee are as follows:

David Wigg (Chair)	Tel.: +44 (0)1732 455446 (H)
	Tel.: +44 (0)207 815 7472 (W)
	E-mail: wiggjd@bcs.org.uk
Monique L'Huillier (Secretary)	Tel.: +44 (0)1276 20488 (H)
	Tel.: +44 (0)1784 443243 (W)
	E-mail: m.l'huillier@rhul.ac.uk
Derek Lewis (Editor)	Tel.: +44 (0)1404 814186 (H)
	Tel.: +44 (0)1392 264296 (W)
	E-mail: d.r.lewis@exeter.ac.uk
Roger Harris (Webmaster)	Tel.: +44 (0)208 800 2903 (H)
	E-mail: rwsh@bcs.org.uk
Douglas Clarke	Tel.: +44 (0)1908 373141
Ian Kelly	Tel.: +44 (0)1276 857599
	E-mail: idkk@idkk.com
Veronica Lawson	Tel.: +44 (0)207 7359060
	E-mail: veronica_1@compuserve.com
Correspondent Members:	
Gareth Evans (Minority Languages)	E-mail: gevans@ubs-wsc.org
Ruslan Mitkov	Tel: +44 (0)1902 322471 (W)
	E-mail: r.mitkov@wlv.ac.uk
Heather Fulford	E-mail: h.fulford@lboro.ac.uk
Mark Stevenson	E-mail: stanage_edge@hotmail.com

BCS Library

Books kindly donated by members are passed to the BCS library at the IEE, Savoy Place, London, WC2R 0BL, UK (tel: +44 (0)207 240 1871; fax: +44 (0)207 497 3557). Members of the BCS may borrow books from this library either in person or by post. All they have to provide is their membership number. The library is open Monday to Friday, 9.00 am - 5.00 pm.

Website

The website address of the BCS-NLTSG is: http://www.bcs.org.uk/siggroup/nalatran

Latin to English Machine Translation - A Direct Approach

by

Paul R. Bowden

Dept. of Computing
The Nottingham Trent University
Burton Street, Nottingham NG1 4BU
E-mail: paul.bowden@ntu.ac.uk

Abstract

The *Brutus* program is a direct-MT attempt at Latin to English translation. Incorporating a Latin morphological processor and two specialised English morphological processors, it produces translations on a sentence by sentence basis in a series of stages. The first stage is a word for word transliteration, each English word being marked with case, gender, tense etc. information inherent in the Latin source word. Subsequent stages correct English noun number and verb endings, find connected noun-adjective groups, mark subject and object, and alter word order to standard English VSO order, amongst other things. An open loop learning mechanism is employed for building the Latin vocabulary. *Brutus* is likely to be useful as a teaching aid or as an assistant to those whose Latin skills are not good. A future possibility is automatic marking of student Latin to English translations.

Introduction

The direct method was the earliest approach to machine translation (MT). Although its deficiencies soon became apparent, it remains popular in certain situations such as uni-directional MT. This is largely due to its usefulness, robustness and relative simplicity. Direct translation involves a series of stages commencing with word-for-word translation, each stage refining the output from the previous stage, e.g. by word-order changes (Hutchins and Somers, 1992).

The Latin language, the tongue of the Roman Empire, began as just one of many languages spoken in pre-Roman Italy, but grew in prominence along with Rome itself (Baugh and Cable, 1978). The Vulgar Latin spoken in the street eventually gave rise to modern Romance languages. Classical Latin, a more highly inflected form, came to be used for literary and cultural purposes, and in later centuries as a means of international communication and storage of knowledge, particularly in the domains of the law, religion and scientific study. It is Classical Latin which is the target of this research (henceforth referred to simply as Latin).

Latin is an inflected language in which sentence word order is used mostly to emphasise certain words rather than to indicate e.g. subject/object/indirect object (SOI) information. In Latin, nouns are inflected to indicate case (subject, object, possessive etc.) and number. Adjectives agree with nouns in number, case and gender. Verbs are conjugated to indicate person, number (singular or plural), voice (active or passive), mood (indicative or subjunctive), and tense. (See e.g. Paterson and Macnaughton (1968) for an introduction to Latin.) This highly inflected nature lends itself well to a direct MT approach. For example, the two-word Latin sentence *Regem laudabo* directly translates as *King I-will-praise,* where the word *regem* is the inflected form of *rex* (king) in the accusative case (used for the direct

object), and where the *-abo* ending of the verb indicates first person singular, future tense of the verb *laudare,* to praise.

The computer program described in this paper attempts Latin to English MT in a multistage direct approach. This program has been named *Brutus.* Although the program is still being developed it is sufficiently advanced to be currently useful. The program incorporates a Latin lexical morphological processor capable of deducing the meaning(s) of a Latin word based upon a stem form. A vocabulary file of already-met words is used for fast look up, and this also holds stem forms available to the morphological processor. The program incorporates an open-loop learning mechanism so that newly deduced word meanings may be presented to a human user for checking before being appended to the vocabulary list.

Direct Look Up and Morphological Processing

The first stage of processing provides a word-for-word translation of the Latin text into English, on a sentence by sentence basis. (Texts of several hundred sentences in length may be processed, so that realistic texts may be handled.) However, the English words so produced are tagged with syntactical and functional information inherent within the original inflected Latin words. Thus for example, the two Latin words given in the example above have the following vocabulary entries:

regem=kings[NOUNMasc-AccSing]
laudabo=I_will_praise[VERB-1stSingFutIndAct]

In the above case we see that the English word *king* is the direct object of the verb (because the Latin was in the accusative case, indicated by *Acc*) and this information is preserved within the square bracketed section for use by later processing stages. (For practical reasons, all noun meanings are stored in their plural English forms; a later processing stage alters the plural form to the singular if necessary.) The word *king* in English is not inflected by functional role; this information in English is instead held by sentence word order. On the other hand, the English *I_will_praise* clearly already holds within it the person and tense information, also stated in the following square bracketed section (the auxiliary *will* is used to indicate future tense in all persons; *shall* is not used, despite its historical use in the 1st person singular and plural). Thus Brutus does not distinguish simple future and emphatic future, traditionally distinguished by swapping *will* for *shall* and vice versa in all person/numbers. Whenever Brutus encounters *laudabo* the above output is looked up in the vocabulary list.

The inflectional nature of Latin means that several tens of forms derived from each stem may need to be held in the vocabulary file. This is in theory not a problem, but in practice requires constant additions of new words for each new text encountered. Without any form of automated assistance the human user would have to create the meaning entry in every case, a time-consuming and tedious task. To counteract this, Brutus contains a morphological processor which derives the meaning entry from the stem form of the new word. Thus the human operator need only check the presented deduced meaning, rather than create it from scratch. The morphological processor does, however, require that the relevant stem form is already in the vocabulary file.

An example is now given. The word *laudabas* is encountered, and is not found in the vocabulary file. However, the vocabulary list is found to contain the entry:

laudare=to_praise[VERB-InfConj1]

This indicates that *laudare* is the infinitive form of the first conjugation verb meaning 'to praise'. The morphological processor finds that *laudabas* may be derived from *laudare* as the active voice, indicative mood, imperfect tense, 2nd person singular form, and suggests the following new vocabulary entry:

laudabas=you_were_praise[VERB-2ndSingImpIndAct]

Here, although *you_were* is inserted by the Latin morphological processor, the stem form of the verb is returned; a later English verb morphology stage alters *praise* to *praising*. The user accepts or rejects this meaning, and the translation continues. Once accepted, the new meaning is added to the vocabulary list and is available to all future runs of the program. Learning mode is selectable so that the program may be run with or without the need for human feedback. The vocabulary file may be filled manually or through the learning mechanism. Brutus is robust and does not stop when an unknown word is encountered; instead, either a deduced meaning is used or if this is not possible (e.g. because the stem form is not in the vocabulary file) the word is left as UNKNOWN. In the rare cases where one Latin word has more than one possible derived meaning, the human user is presented with all meanings in turn for acceptance or rejection. Likewise, when translation is proceeding normally using look up, Brutus returns all looked-up meaning forms for each Latin word in the text, and these are disambiguated later.

Further Processing Stages

To illustrate the intended processing stages following the initial direct stage, the following example sentence will be used:

Secunda legio castra in Gallia habet, sed in Britanniam cum imperatore festinabit.

STAGE 1 The output from the first stage is as follows:

Second[ADJ-NomSingFem,VocSingFem,AblSingFem,NomPlurNeut,VocPlurNeut,AccPlurNeut] **legions**[NOUNFEm-NomSing,VocSing] **camps**[NOUNNeut-NomPlur,VocPlur,AccPlur] **in**[PREP-+Abl]-OR-**into**[PREP-+Acc] **Gaul**[NOUNFem-NomSing,VocSing,AblSing] **he/she/it_have**[VERB-3rdSingPresIndAct] , **but**[CONJ] **in**[PREP-+Abl]-OR-**into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **generals**[NOUNMasc-AblSing] **he/she/it_will_hurry**[VERB-3rdSingFutIndAct] .

Nouns, with the exception of proper nouns, are given in their plural form at this point, as stored in the vocabulary. Verbs are in the form of pronoun plus modals indicating tense/number plus uninflected stem (e.g. he/she/it_eat, I_will_praise, you_had_eat). The forward slash is used to separate alternatives within entries, but where the Latin word has more than one quite distinct sense the possible senses are each given using *-OR-* as the separator (see e.g. for *in*).

STAGE 2 This stage corrects the noun numbers where possible and also morphs verb forms to correct English. The output from the second stage is as follows:

Second[ADJ-NomSingFem,VocSingFem,AblSingFem,NomPlurNeut,VocPlurNeut,AccPlurNeut] **legion**[NOUNFEm-NomSing,VocSing] **camps**[NOUNNeut-NomPlur,VocPlur,AccPlur] **in**[PREP-+Abl]-OR-**into**[PREP-+Acc] **Gaul**[NOUNFem-NomSing,VocSing,AblSing] **he/she/it_has**[VERB-3rdSingPresIndAct] , **but**[CONJ] **in**[PREP-+Abl]-OR-**into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **general**[NOUNMasc-AblSing] **he/she/it_will_hurry**[VERB-3rdSingFutIndAct] .

Here, *legions* and *generals* have been converted to their singular forms, since the string 'Plur' does not occur in the square bracketed sections following them. This is achieved using the *sing* function described in Bowden, Halstead and Rose (1996). Also in this stage the stem verb form

have has been altered to *has*. This is done using a look-up table containing rules for the construction of irregular verb forms, such as those shown in Table 1. Stem/[...] pairs not present in the table are altered in a regular manner by a rule-based system which examines the stem ending, although in many cases the stem need not be altered (see e.g. the *laudabo* example above).

Form	Result	Example
have[3rdSingPresIndAct]	has	he/she/it_have → he/she/it_has
eat[...PerfIndAct]	ate	I_eat → I_ate
drink[...PerfIndAct]	drank	you_drink → you_drank
drink[...PlupIndAct]	drunk	we_had_drink → we_had_drunk
bite[...PlupIndAct]	bitten	they_had_bite → they_had_bitten
teach[...PerfIndAct]	taught	I_teach → I_taught
get[...FutperfIndAct]	got	we_will_have_get → we_will_have_got

Table 1. Irregular English verb morphological processing rules

This morphological processing is specifically tailored to the task at hand; the rules are triggered by the Latin tenses etc. as found in the square bracketed parts. Brutus always translates the Imperfect tense as a continuous past (e.g. we were eating) and always translates the Perfect as a simple past (e.g. I ate). This is not ideal, but it is pragmatic. The other Latin tenses present no such dilemmas.

STAGE 3 This stage looks for connected adjective and noun runs, and reduces the possibilities in the square brackets accordingly.

Adjectives in Latin may follow or precede the noun, and in addition there is a construct where *et* (and) is used to link two adjectives applying to the same noun. The output is as follows:

> **Second**[ADJ-NomSingFem,VocSingFem] **legion**[NOUNFEm-NomSing,VocSing] **camps**[NOUNNeut-NomPlur,VocPlur,AccPlur] **in**[PREP-+Abl]-OR-**into**[PREP-+Acc] **Gaul**[NOUNFem-NomSing,VocSing,AblSing] **he/she/it_has**[VERB-3rdSingPresIndAct] , **but**[CONJ] **in**[PREP-+Abl]-OR-**into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **general**[NOUNMasc-AblSing] **he/she/it_will_hurry**[VERB-3rdSingFutIndAct] .

In this case the plural possibilities have been removed from *Second* since the noun it modifies, *legion*, is in the singular, and also the Abl possibility has been deleted because it is not there for the noun. The difficulty in this stage lies with finding the boundaries between adjective-noun groups; for example, the noun *camps* is not part of the first group. This is done using heuristics which tell where to place the boundary for each possible ADJ/NOUN run.

STAGE 4 This stage resolves prepositions, both internally (/-parts) and between polysemous forms (-OR- parts). The output becomes:

> **Second**[ADJ-NomSingFem,VocSingFem] **legion**[NOUNFEm-NomSing,VocSing] **camps**[NOUNNeut-NomPlur,VocPlur,AccPlur] **in**[PREP-+Abl] **Gaul**[NOUNFem-AblSing] **he/she/it_has**[VERB-3rdSingPresIndAct] , **but**[CONJ] **into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **general**[NOUNMasc-AblSing] **he/she/it_will_hurry**[VERB-3rdSingFutIndAct] .

Here, the first occurrence of Latin *in* can only mean 'existing inside' (Latin *in+Abl* can sometimes mean 'on', but Brutus always translates it as 'in'; a future stage will perform in-on changing where necessary) since what follows is possibly *Abl* but definitely not *Acc*. The second occurrence of Latin *in* must mean 'into', since what follows is an *Acc* noun. Redundant case

information is deleted from within the [...] parts. The important factor is the case of the following noun phrase, but this example shows that a list of place names is also maintained for /-resolution.

STAGE 5 This stage marks the Subject, Verb, Object, and Everything-Else parts (SVOE structure). This is done purely by part of speech and case as indicated with the square bracketed parts:

> <Subj1=**Second**[ADJ-NomSingFem] **legion**[NOUNFEm-NomSing]> <Obj1=**camps**[NOUNNeut-AccPlur]> <Else1= **in**[PREP-+Abl] **Gaul**[NOUNFem-AblSing]> <Verb1=**he/she/it_has**[VERB-3rdSingPresIndAct]> , <Else2=**but**[CONJ] **into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **general**[NOUNMasc-AblSing]> <Verb2= **he/she/it_will_hurry**[VERB-3rdSingFutIndAct]> .

Angle brackets are used to indicate the Subject etc parts as illustrated. It is assumed that all sentences have a subject. In the example, this allows various cases of nouns to be deleted from within the square brackets. Subjects are identified first so that nouns which might be either subject or object can be disambiguated.

STAGE 6 Following the identification of possible Subject, Object etc entities in Stage 5, this stage links verbs to their subjects and deletes /-parts within the verbs:

> <Subj1(Verb1,Verb2)=**Second**[ADJ-NomSingFem] **legion**[NOUNFEm-NomSing]> <Obj1(Verb1)=**camps**[NOUNNeut-NomPlur,AccPlur]> <Else1= **in**[PREP-+Abl] **Gaul**[NOUNFem-AblSing]> <Verb1=**has**[VERB-3rdSingPresIndAct]> , <Else2=**but**[CONJ] **into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **general**[NOUNMasc-AblSing]> <Verb2= **will_hurry**[VERB-3rdSingFutIndAct]> .

Here, the first verb (has) is linked to the subject, which has Sing marking. The second verb (he/she/it_will_hurry) might have posed more of a problem, since it might have applied to either the *legion* in the first clause or to the *general* in the second. However, *general* is not marked as a subject group, and so this possibility may be discounted; the second verb must be linked to the single subject of the sentence.

STAGE 7 In this stage, word order is altered to reflect standard English SVOE order and <...> parts can therefore be removed. This results in the following:

> **Second**[ADJ-NomSingFem] **legion**[NOUNFEm-NomSing] **has**[VERB-3rdSingPresIndAct] **camps**[NOUNNeut-AccPlur] **in**[PREP-+Abl] **Gaul**[NOUNFem-AblSing], **but**[CONJ] **will_hurry**[VERB-3rdSingFutIndAct] **into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **general**[NOUNMasc-AblSing] .

Moving the verb positions also allows removal of he/she/it where the subject immediately precedes the verb or where it is deemed to be ellipted from that position. In the case of a second verb, where one verb has already been linked to a subject and moved, the second verb is moved to the start of the clause it appears in (after a conjunction if one is present).

STAGE 8 The translation is nearly complete. This penultimate stage cannot be performed on an isolated sentence, for it comprises the insertion of determiners. (In Latin, the specificity of objects within a text is deduced largely by the reader, rather than being explicitly marked as it is in English.) The entire translated text after stage 7 will be passed to a discourse-entity (DE) recogniser, which will use the [...] parts to detect and count each possible DE and insert definite

and indefinite articles. Assuming that the example sentence is mid-way through a longer text, this would result in the following:

> **The second**[ADJ-NomSingFem] **legion**[NOUNFEm-NomSing] **has**[VERB-3rdSingPresIndAct] **camps**[NOUNNeut-AccPlur] **in**[PREP-+Abl] **Gaul**[NOUNFem-AblSing], **but**[CONJ] **will_hurry**[VERB-3rdSingFutIndAct] **into**[PREP-+Acc] **Britain**[NOUNFem-AccSing] **with**[PREP-+Abl] **the general**[NOUNMasc-AblSing] .

STAGE 9 In the final stage, all square-bracket text is removed, and underscores replaced by spaces:

The second legion has camps in Gaul, but will hurry into Britain with the general.

Secunda legio castra in Gallia habet, sed in Britanniam cum imperatore festinabit.

Discussion

The Brutus program is still being developed and although coding for stages 1, 2 and 9 is complete, much of the remaining stages is still being built. In addition, much more vocabulary needs to be added to the program's dictionary.

The above description does not discuss certain tasks such as the linking of adverbs to verbs. There are also constructions in Latin which take a standard form (e.g. the 'ablative absolute' construction, the use of *-ne* on the first word of a sentence to indicate a question etc). These will be tackled within the above stages or in separate stages at relevant positions within the above framework.

At this early stage, it is difficult to know how much of a problem polysemous words will be. It is thought that these are in fact much rarer than in English, due to the highly inflected nature of Latin. It is possible to concoct examples where one orthographic Latin word has have more than one distinct meaning, but it remains to be seen if this is a problem in practice.

Brutus embodies a very shallow approach to MT, but even the output from Stage 1 (completed) is largely understandable and in itself is a good translation aid. Thus Brutus is already useful. The question arises as to who might use such a program; clearly, expert classical scholars are unlikely to need such assistance. However, as a teaching aid Brutus might well prove very helpful. Students could use it to check their translations, and in this sense it might actually be more useful than a human marker, since human markers do not usually write down e.g. all the possible case/gender combinations for each noun in the text. Also, as the basis for an automatic marking system, Brutus has the potential to do more than just highlight incorrectly translated words or phrases, since it could in effect *explain* the correct translation to the student.

References

Baugh, A. C. and Cable, T. (1978) *A History of the English Language* (3rd Edition) Routledge and Kegan Paul

Bowden, P. R., Halstead, P. and Rose, T. G. (1996) *Dictionaryless English Plural Noun Singularisation Using A Corpus-Based List of Irregular Forms* In *Corpus-based Studies in English - Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)* Stockholm, May 15 - 19 1996, Rodopi

Hutchins, W. J. and Somers, H. L. (1992) *An Introduction to Machine Translation* Academic Press

Paterson, J. and Macnaughton, E. G. (1968) *The Approach to Latin (First Part)* (revised 1968) Oliver and Boyd

D:\WINNT\Profiles\cmp3bowdepr\Personal\Latin.doc

Learning Machine Translation Strategies Using Commercial Systems: Discovering Word Reordering Rules

by

Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant, Spain.
E-mail: **mlf@dlsi.ua.es**

Abstract

Commercial machine translation (MT) systems seldom publicize their MT strategies; however, unlike well-documented experimental systems, they are easily available, some even as a free Internet service. Some of these systems seem to follow very closely a *basic model* (a sort of advanced morphological transfer strategy), described in detail in this paper, which focuses on the translation from English to Spanish, and, in particular, on the mechanisms used by the systems to obtain the correct word order for the target language. The *basic model* is so simple that a laboratory assignment based on it allows students to discover interesting details about the operation of a number of real MT systems, such as the reordering rules used.

Introduction

Many universities teach undergraduate and graduate courses dealing with the subject of machine translation (MT); part of these courses is expectedly devoted to teaching MT strategies. On the other hand, commercial MT systems are readily available, either as low-priced software packages for PCs (ranging from 30 to 300 €) or as free Internet servers. A recent survey (Balkan et al. 1997) has dealt, among other aspects, with the use of commercial MT systems in teaching; while almost all respondents agree that 'while using a working MT system to teach MT is definitely beneficial, it involves a huge amount of work'. The authors 'had hoped to find […] that someone had done all the hard work' but got negative results. The recent increase in Internet availability of commercial MT systems may alleviate another problem described in the study, namely, that '[…] those interested in obtaining an MT system said they were prepared to invest very little' (a few hundred € and from one day to one week). The work reported in this paper covers ways to use readily available commercial systems in teaching MT strategies. I do not claim to have done 'all the hard work', but expect this proposal to be useful for the community of MT instructors.

In particular assignments presented here, the source language (SL) is English and the target language (TL) is Spanish, because most of our translation and computer science students are familiar with this pair.

The hypotheses are compatible with a transfer MT architecture (Arnold et al. 1994, Arnold 1993; Hutchins and Somers 1992) called here the *basic model* (see above). The basic model explains, at least partly, the behavior of the three systems covered: Globalink's *Power Translator Pro 5.0* or *Power Translator Deluxe* (PT), which are basically equivalent, Transparent Technologies' *TranscendRT* (TRT), and Softissimo's *Reverso*.

A note of caution is necessary: the models proposed are derived from a black-box study of these systems, in the absence of documentation by the manufacturers; the models may

therefore be partially incorrect or inaccurate, but this does not invalidate their use in the laboratory as long as they explain the basic behaviors observed and solve the question 'Why do I get this *word salad?*'. In particular, the models may be used to stress the mechanical or rule-based nature of machine translation in front of frequent student misconceptions about the behavior of computers, particularly among non-computer-science majors. Indeed, in one case, the rule set induced was completely confirmed by the contents of one of the system files (Mira i Gimènez and Forcada, 1998). While some manufacturers claim that their products actually perform syntax analysis (http://www.transparentlanguage.com/ets/about/transcendrt.htm) close observation reveals behaviors that are not compatible with what most experts would call syntax analysis (for example, the ability to identify and correctly process simple constituents - such as noun phrases - regardless of their length).

The 'basic model'

The basics of the behavior of the commercial MT systems examined may be conveniently explained using a *basic model*, a simple transfer architecture (Hutchins and Somers, 1992; Arnold et al., 1994; Arnold, 1993) which is more advanced than a morphological transfer system but cannot be properly called a syntactic transfer system because it does not perform full syntax analysis (the *basic model* may also be considered as a special case of what Arnold et al. (Arnold, 1993, section. 4.2) call a *transformer architecture* or what Hutchins and Somers (Hutchins and Somers, 1992, section 4.2) refer to as a *direct* architecture). We have found this *basic model* to be simple enough for students to understand it and apply it to explain and predict the behavior of an MT system. The model is neither aimed at describing MT systems in general nor to encompass the basic problems of MT, but rather to describe in simple terms the behavior of a set of real, commercially available MT systems.

Four basic tasks are clearly distinguishable in the basic model: *morphological analysis*, which yields all possible lexical forms (LFs) for each surface form (SF) in the text (in particular, morphological analysis takes each SF or word (e.g., *taught*) and builds one or more LFs per word, consisting of a lemma or canonical form (*teach*), the lexical category (*verb*), and inflection information (*past tense*)); *homograph disambiguation* (a homograph is a SF having more than one LF), which chooses one of the LFs, usually using simple rules based on the lexical categories assigned to neigboring words; the *transfer* task itself (see below); and *morphological generation*, which transforms each of the TL lexical forms (TLLFs) into the corresponding SFs.

The transfer task is organized around *patterns* representing fixed-length sequences of source-language LFs (SLLFs); two sequences are equivalent if they contain the same sequence of lexical categories. The system contains a catalog of the patterns it knows how to process. Patterns are not 'phrases' or constituents in the syntactic sense, because they are flat and unstructured, but pattern detection is an advancement with respect to bare morphological analysis and may be considered as a rudimentary form of syntax analysis.

The *pattern detection* phase occurs as follows: if the transfer module starts to process the *i*-th SLLF of the text, l_i, it tries to match the sequence of SLLFs l_i, l_{i+1},... with all of the patterns in its pattern catalog: the longest matching pattern is chosen, the matching sequence is processed (see below), and processing continues at SLLF l_{i+k}, where k is the length of the pattern just processed. If no pattern matches the sequence starting at SLLF l_i, it is translated as an isolated word and processing restarts at SLLF l_{i+1} (when no patterns are applicable, the

systems resort to word-for-word translation). Note that each SLLF is processed only once: patterns do not overlap; hence, processing occurs left to right and in distinct 'chunks'.

Pattern processing takes the detected sequence of SLLFs and builds (using the bilingual dictionary) a sequence of TLLFs which may be completely reordered, with LFs added to it or deleted from it. The inflection information in TLLFs is generated so that agreement is observed inside the sequence if necessary. For instance, the English pattern article-adjective-noun (such as 'the red tables') is turned into the Spanish sequence article-noun-adjective ('las mesas rojas'), after propagating the gender and number of 'mesas' to both the article and the adjective. In addition, the transfer module may maintain 'state' information to ensure left-to-right interpattern relationships such as subject-verb number agreement. State information may be updated after each pattern is processed.

A finite catalog of fixed-length 'frozen' sequences cannot possibly cover all of the possible forms a certain constituent (i.e., a noun phrase) may take, because of recursivity in grammar rules (for example, there is no theoretical limit to the number of noun phrases in the possessive case inside a given noun phrase). However comprehensive the catalog is (consider also that the number of patterns grows dramatically with length), the system will always find unknown phrases: a pattern will match part of the constituent, process it as a complete constituent, and leave the trailing words up for further processing; this usually results in a 'word salad', which would be very hard to interpret; however, having the basic model in mind, these garbled translations give invaluable cues as to which are the particular patterns in the system's catalog; this will be exploited in the laboratory assignment proposed.

Laboratory assignment

The purpose of this laboratory assignment is to discover the reordering patterns used by the three MT systems studied. It is designed for a two- or three-hour session (but may be cut down by reducing the number of machine translation programs studied), and requires substantial guidance by the course instructor. This assignment should be placed after a preliminary assignment in which a comparison of the translation of a set of sentences proposed by the instructor with the translation of each word in isolation --one word per line, with a blank line between words-- reveals processes such as context-dependent homograph disambiguation (part-of-speech tagging), use of multiword units, *word reordering*, and agreement enforcement.

In the assignment, students will be asked to study the behavior of PT, TRT and Reverso when translating noun phrases of growing complexity from English to Spanish, to try to understand the strategy they use. Initially, they will study the translation of the 10 sentences represented by the expression *I saw the [[senior] [computer] expert's] [large] desk*. An acceptable translation, resulting from considerable reordering of words, is *Vi el escritorio [grande] [del experto [de computadora] [mayor]]*.

Students will be told to assume that the system does not perform real syntactical analysis, but instead uses the strategy explained in this paper. For simplicity, they will be recommended not to consider articles as part of the patterns as a first approximation.

Students will write down, for each English sentence, the translation produced by each program, the nearest acceptable Spanish translation, and, where differences are significant, a possible explanation in terms of the proposed strategy. The following questions may be used to guide their work: 'Can you identify parts of the sentence which have been independently

processed? Which are the active patterns in each program? Why do we get incorrect translations for some sentences?'

If time allows, students will be invited to confirm the details of their hypothesis (patterns, etc.) with more noun phrases having the same sequences of lexical categories but different words (adjectives, nouns and nouns in the possessive case), or different composition. Make sure they do not introduce any homograph.

Hints for the instructor

What follows is an analysis of the results produced by each of the programs, to help the instructor guide the students during the assignment.

Power Translator Pro 5.0 (PT):

1. *I saw the desk* → *Yo vi el escritorio*: Acceptable. No reordering occurs.
2. *I saw the large desk* → *Yo vi el escritorio grande*: Acceptable. A reordering occurs in *large desk*, which may be explained with rule $R_1 : A\ N \rightarrow N\ A$ where *N* stands for a noun and *A* stands for an adjective.
3. *I saw the expert's desk* → *Yo vi escritorio del experto*: Acceptable except for articles. The reordering in *expert's desk* may be explained with a new rule: $R_2 : NG_1\ N_2 \rightarrow N_2$ **d** N_1, where *NG* stands for a noun in the possessive (genitive) case and **d** for the preposition *de*.
4. *I saw the expert's large desk* → *Yo vi el escritorio grande de experto*: Acceptable except for articles. The reordering in *expert's large desk* has to be explained with a new rule ($R_3 : NG_1\ A\ N_2 \rightarrow N_2\ A$ **d** N_1), because R_1 would have yielded **experto escritorio grande*, and R_2 cannot be applied.
5. *I saw the computer expert's desk* → *Yo vi el escritorio de experto de computadora*: Acceptable except for articles. The reordering in *computer expert's desk* has to be explained with a new rule ($R_4 : N_1\ NG_2\ N_3 \rightarrow N_3$ **d** N_2 **d** N_1), because only R_2 may be applied and it would have yielded **computadora escritorio de experto*.
6. *I saw the computer expert's large desk* →**Yo vi la computadora escritorio grande de experto*. Unacceptable: PT splits the noun phrase *computer expert's large desk*. Only *expert's large desk* is reordered, using rule R_3, because PT has no rule matching the sequence *N NG A N*.
7. *I saw the senior expert's desk* → *Yo vi el escritorio de experto mayor*: Acceptable except for articles. The reordering in *senior expert's desk*, that is, in a sequence *A NG N*, has to be explained with a new rule, $R_5 : A\ NG_1\ N_2 \rightarrow N_2$ **d** $N_1\ A$, because only R_2 could have been applied, with the result **mayor escritorio de experto*.
8. *I saw the senior expert's large desk* → **Yo vi el mayor escritorio grande de experto*. Unacceptable: PT splits the noun phrase *senior expert's large desk*. Once again, only *expert's large desk* is reordered, using rule R_3. PT's catalog does not contain the pattern *A NG A N*.

9. *I saw the senior computer expert's desk* → **Yo vi la computadora mayor escritorio de experto*. Unacceptable: PT splits the noun phrase *senior expert's large desk*. First, rule R_1 is applied to *senior computer* and then rule R_2 is applied to *expert's desk*. PT's catalog does not contain the pattern *A N NG N*.
10. *I saw the senior computer expert's large desk* → **Yo vi la computadora mayor escritorio grande de experto*. Unacceptable: PT splits the noun phrase *senior computer expert's large desk*. First, rule R_1 is applied to *senior computer*, and then rule R_3 is applied to *expert's large desk*. PT's catalog does not contain the pattern *A N NG A N*.

Optional work with PT:

As was explained in detail in (Mira i Gimènez and Forcada, 1998), PT stores the patterns in an ASCII text file, named `engspan.pat` in directory `dicts`. An extended assignment may involve examining the file, or even modifying it to change PT's behavior.

TranscendRT (TRT):

TRT's strategy is analogous to that of PT, but with some differences in rules: R_1 is applied to sentences #2, #6, and #10; R_2 is applied to sentences #3 and #6 (in #6, TRT applies the rule ignoring the possessive case in N_2); R_3 is used in #4; R_4 is used in sentence #5. The new rules are:

R'_5:

$A\ NG_1\ N_2 \rightarrow N_3\ A\ \mathbf{d}\ N_1$ (the rule assumes that the adjective modifies the second noun; used in #7 and #10; in #10, TRT ignores the possessive case in N_2).

Comment [1]: MATH $A\; NG_1\; N_2\; \rightarrow N_2\; A\; {\bf d}\; N_1$

R'_6:

$A_1\ NG_1\ A_2\ N_2 \rightarrow N_2\ A_1\ A_2\ \mathbf{d}\ N_1$ (it reorders adjectives inadequately in this case but may be correct in other cases; used in #8).

Comment [2]: MATH $A_1\; NG_1\; A_2\; N_2\; \rightarrow N_2\; A_1\; A_2\; {\bf d}\; N_1$

R'_7:

$A\ N_1\ NG_2\ N_3 \rightarrow N_3\ A\ \mathbf{d}\ N_2\ \mathbf{d}\ N_1$ (it assumes that the adjective affects the third noun; used in #9).

Comment [3]: MATH $A\; N_1\; NG_2\; N_3\; \rightarrow N_3\; A\; {\bf d}\; N_2\; {\bf d}\; N_1$

In addition, TRT deletes the pronoun *Yo* before the verb *vi*, translates *expert* as *perito* instead of *experto* and has a different treatment for articles.

Reverso:
Reverso's strategy is very similar, with some differences in the rules applied. The main difference is the addition of the Spanish preposition *a*, mandatory before direct-object noun phrases having a person as their head. Rules R_1-R_4 are as in TRT, and R_5 as in PT. The new rules are:

R_6':

New rule: $A_1\ NG_1\ A_2\ N_2 \rightarrow N_2\ A_2\ \mathbf{d}\ N_1\ A_1$, used in #8.

R_7':

New rule: $A_1\ N_1\ NG_2\ A_2\ N_3 \rightarrow N_3\ A_2\ \mathbf{d}\ N_2\ \mathbf{d}\ N_1\ A_1$, used in #9.

Concluding remarks

I have shown how a very simple model of MT may be very useful to obtain a rather detailed explanation --including the formulation of rules-- for the word-reordering behavior of three readily available commercial MT systems, and how this work may be organized as a laboratory assignment in which the students use the model to formulate the particular rules used by each system under the guidance of the laboratory instructor.

Bibliography

Arnold, D. (1993) 'Sur la conception du transfert', in Bouillon, P. and Clas, A., editors, *La traductique*, pages 64-76. Presses Univ. Montréal, Montréal

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., and Sadler, L. (1994) *Machine Translation: An Introductory Guide*. NCC Blackwell, Oxford. Available as `http://clwww.essex.ac.uk/~doug/MTbook/`

Balkan, L., Arnold, D., and Sadler, L. (1997) *Tools and techniques for machine translation teaching: a survey*. Technical report, University of Essex, Colchester, Essex, U.K. URL: `http://clwww.essex.ac.uk/group/projects/MTforTeaching`

Hutchins, W. and Somers, H. (1992) *An Introduction to Machine Translation*. Academic Press

Mira i Gimènez, M. and Forcada, M. L. (1998) Understanding PC-based machine translation systems for evaluation, teaching and reverse engineering: the treatment of noun phrases in Power Translator, in *Machine Translation Review (British Computer Society)*, 7:20-27. (available at `http://www.dlsi.ua.es/~mlf/mtr98.ps.Z`)

Note

This work has been supported by the Spanish Comisión Interministerial de Ciencia y Tecnología through grant TIC97-0941 and also by the Caja de Ahorros del Mediterráneo.

An Example Based MT System in News Items Domain from English to Indian Languages

by

Dr. Sivaji Bandyopadhyay
Jadavpur University, Calcutta, India

Abstract

The paper reports an on-going research project on a Knowledge Driven Generalized Example-Based Machine Translation system from English to Indian languages. It is currently translating short single paragraph news items from English to Bengali. Headlines are translated using knowledge bases and example structures. The sentences in the news body are translated by analysis and synthesis. Semantic categories are associated with words to identify the inflections to be attached in the target language and to identify the context in the sentence. Context identification is also done by using context templates for each word. The example base also includes the mapping of grammatical phrases from the source to the target language. The methodologies can be used for developing similar systems for other Indian languages.

Machine Translation in the News Items Domain

The domain of news items has attracted the attention of Machine Translation (MT) researchers all over the world. In India, a human-aided MT system for translating English news texts to Hindi is being developed at the National Centre for Software Technology, Mumbai (http://www.ncst.ernet.in/kbcs/NLP.html).

The MT Research Group at the Information Sciences Institute (ISI), University of Southern California (USC), is developing programs that translate Japanese, Arabic and Spanish unrestricted newspaper texts into English. The work is reported in Knight et al (1994a), Hatzivassiloglou (1995) and Yamada(1996).

The Pangloss Mark III machine translation system translates unrestricted Spanish news-wire text into English. The work was carried out : Center for Machine Translation, Carnegie Mellon University (CMT/CMU), ISI/USC and Computing Research Laboratory, New Mexico State University (CRL/NMSU). The translation from Russian News items into English was done at the CRL/NMSU. Details for this system can be obtained from Knight(1994b) and Brown(1996).

The NHK System in Japan which translates English newspaper articles to Japanese is described in Hutchins (1999). The improvement of translation quality of English Newspaper headlines by automatic pre-editing, in the English to Japanese machine translation system being developed at the Sharp Corporation of Japan, is discussed in Yoshimi (1999).

The translation of names and technical terms is crucial in translating news items since these are not found in bilingual dictionaries. The results and examples of transliteration rules for names and technical terms from Arabic to English can be found in Knight(1997) and Stalls(1998).

Proper nouns like the place of news, month, date and the name of the news agency can be recognized since they occupy fixed places in a news item. The capitalization of proper nouns has been assumed in the present work. The Bilingual dictionaries include the proper noun, its

classification and the target language representation. Proper nouns may also arise in a news item as acronyms. A separate Acronym Dictionary is used in the present system.

News Items : Structure & Classification

Short single paragraph news items are being sampled from the CALCUTTA Edition of the English News paper *The Statesman*. Generally these samples follow the structure :

< Headline >

<Place of News>, <Month>.<Date>. - <Body of the News> - <News Agency>.

The news items are classified based on the nature of the news. The different acronyms and proper nouns may vary for each classification and separate Bilingual dictionaries are kept for each of them. If the Example bases and the Knowledge bases are classified then the searching into them is focused, too.

Translation of the News Headlines

Translation of news headlines is being carried out using Example-based MT techniques. The Example base includes both specific and general examples. It includes general examples at the syntactic level also. The various phrases in the source language and their corresponding translation in the target language are stored. The present system thus can be termed as a Generalized Example based Machine Translation system.

The news headlines can be single word (e.g., *Repolling*), multiple word (e.g., *Cease fire plea*), an ungrammatical sentence generally without the auxiliary verb (e.g., *Kanika critical*) or a grammatical sentence (*Aatre is new Scientific Advisor*). The translation for the headline is first searched in the table, organized under each headline structure, containing specific source and target language pairs. If not found, the template representation of the headline is searched in another table. Headlines of similar structure are grouped into templates. For example, the two headlines *'Milk to be costlier in Punjab''* and *'Bread to be dearer in State''* follow the template structure : *<item> to be <costlier> in <place>*. The Bengali translation is stored as : *<place>-e <item>-er daam baarlo*.

If the headline still can not be translated, syntax directed translation techniques are applied. If it matches with any phrase of a sentence structure the translation is obtained using the Example base and the Bilingual dictionaries. Otherwise, word-by-word translation is attempted.

Dictionary Design

Root words in English are generally stored in a Dictionary. The *category* (e.g., proper, common etc.), *number* and *gender* are stored for a *noun word* along with its *semantic category*. Semantic categories are associated with words to identify the inflections to be attached with corresponding words in Indian languages as well as to identify the context in the sentence. The semantic category can be independent of the application domain, e.g., Tiger : *<Animate Object>* & Book : *<Inanimate Object>*. Some semantic categories can be dependent on the application domain, e.g., Microsoft : *<Company Name>*. The word *share* has different meanings in *Equal share* and *Microsoft share.* The *person*, *case*, *number* and

gender are stored for *pronoun words*. For words of all other categories only the part of speech information is stored.

The design of the Bilingual Dictionary requires an understanding of the context in which a source language word is used. *Context templates* have been associated with words to identify the context in which the word is used so that its appropriate meaning in the target language can be retrieved from the Bilingual Dictionary. The meaning of a word in English may vary under different parts of speech. Further, the meaning of a word in English may be independent of the context (e.g. *boy*), may depend on the occurrence of a sequence of words (e.g., *run across*) or words with certain semantic categories (e.g., *run a <fever>* and the words *fever* & *temperature* associated with the semantic category *<fever>*) or may depend on the occurrence of certain keywords or keywords with certain semantic categories (e.g. *bank* along with the keyword *river* or *bank* along with keywords with semantic category *<Financial Institution>* have different meaning). These context templates are included in the Bilingual Co-occurrence Dictionaries along with appropriate pointers from the main Bilingual Dictionary which contains the root words in the source language and their meanings.

Context identification is also done by the recognition of Figure of Speech expressions. A separate Figure of Speech Dictionary stores such expressions in English along with their corresponding counterparts in the target language.

Different phases of the Translation System

The work is being carried out in the Visual Studio 6.0 environment with Visual C++ 6.0 as the programming language and Microsoft Access 2000 as the associated database management system. This section describes the different phases of the syntax directed translation in the present system. The general structure of a simple assertive sentence in English in active voice is represented as follows:

{Adverb | Preposition} {Noun Phrase1} {Verb Group} {Adverb | Preposition}{Noun Phrase2}.

Other types of simple sentences are not possible in news items domain. Complex sentences are translated by identifying the *main clause* and the *dependent clause* and then translating the two clauses separately. Similarly, the compound sentences are translated by identifying the conjunction or disjunction in the sentence and translating the two parts separately.

The Knowledge bases include the Suffix Table for Morphological Analysis of English surface level words, Parsing table for Syntactic Analysis of English, Bilingual Dictionaries for different classes of proper nouns, different dictionaries, different tables for synthesis in the target language. The Morphological Analysis is done using Suffix Tables and the English Dictionary. The Suffix Table stores the *suffixes in reverse order*, the *original end pattern* and the *changed end pattern* of the word. The Syntactic Analysis is done using a table driven parser. Auxiliary verbs in English, which do not directly translate into Indian languages but help in identifying the *tense* of the verb are represented as *null* after the analysis phase unless it is the main verb of the sentence. The tense information for a verb is obtained during suffix analysis. Special verb forms like *went* are stored in a separate table and no suffix analysis is required for them. Similar tables are present for irregular noun forms like *men* etc.. The *number* and *person* information for a verb are obtained from the immediately preceding noun or pronoun. Analysis of adjectives takes into account words like *more, most, less, least* etc. occurring before it. Idiomatic expressions are taken care of separately.

The root words in the target language are retrieved. Each surface level word is then synthesized. Each phrase in the input sentence is now considered and its corresponding mapping in the target language is retrieved from the Example base. Special consideration have been made for prepositional phrases. Some inflections are added to the last noun in the prepositional phrase based on the matra of the last symbol of the word or on the semantic category of the word. For example, *on Monday* is translated as *sombaare* but *on the table* is translated as *tebiler upar*. The semantic category of the word *Monday* is <*day*>, its translation to Bengali is *sombaar* and the inflection that has been added to the word is *–e*. The semantic category of the word *table* is neither *date* nor *day*, its translation to Bengali is *tebil* and the inflection that has been added is *–er upar*.

Finally, the translation of the different phrases are assembled. The simple assertive sentences with multiple phrases are translated from English to Indian languages by the following rule :

- the order of all the phrases before the verb phrase are inverted.
- the order of all the phrases after the verb phrase are similarly inverted.
- the verb phrase is put at the end of the sentence.

For example, the different phrases of the sentence '*He has been cultivating the land with a spade in the garden since morning for better crop.*' and their translation into Bengali are as follows :

He – Noun Phrase – *Se*

has been cultivating - Verb Phrase – *chaas karchhilo*

the land – Noun Phrase – *jami*

with a spade – Prepositional Phrase (PP) – *kodaal diye*

in the garden – PP – *baagaane*

since morning - PP – *sakaal theke*

for better crop - PP – *bhaalo fasoler janya*

The Bengali translation of the sentence is as follows :

Se bhaalo fasoler janya sakaal theke baagaane kodaal diye jami chaas karchhilo.

Results and Discussion

The methodologies for a Knowledge driven Generalized Example based Machine Translation system from English to Indian languages in the news items domain have been developed. Currently, a prototype of the system in English-Bengali translation has been developed. The post-editing part of the system is not yet ready. The methodologies can be used for developing similar systems for other Indian languages.

Acknowledgements

The work is being carried out as part of a Research Award granted to the author by the University Grants Commission (UGC), Government of India in 1999 (UGC Research Award for the IXth Plan Period), F.30-95/98 (SA-III).

References

Brown Ralf D. (1996) 'Example-Based Machine Translation in the Pangloss System', in *Proceedings of the COLING-96*

Hatzivassiloglou V. and K. Knight (1995) 'Unification-Based Glossing', in *Proceedings of the 14th IJCAI Conference*

Hutchins J. (1999) 'The Development & Use of Machine Translation Systems and Computer-based translation tools', in the *International Symposium on Machine Translation and Computer Language Information Processing*

Knight K. et al (1994a) 'Integrating Knowledge Bases and Statistics in MT', in *Proceedings of the 1st AMTA Conference*

Knight K. and S. Luk (1994b) 'Building a Large-Scale Knowledge Base for Machine Translation', in *Proceedings of the AAAI-94*

Knight K. and J. Graehl (1997) 'Machine Transliteration' in *Proceedings of the ACL-97*

Stalls B. and K. Knight (1998) 'Translating Names and Technical Terms in Arabic Text', in *COLING/ACL Workshop on Computational Approaches to Semitic Languages*

Yamada K. (1996) 'A Controlled Skip Parser', in *Proceedings of 2nd AMTA Conference*

Yoshimi T. (1999) 'Improvement of Translation Quality of English Newspaper Headlines by Automatic Preediting', in *Proceedings of the MT Summit VII*

Towards a Fully-Automatic High-Quality Machine Translation System for Unrestricted Text

by

Dr Mirosław Gajer
Department of Automatic Control
Technical University in Cracow
al. Mickiewicza 30
30-059 Kraków
Poland
mgajer@ia.agh.edu.pl

Introduction

Machine translation is a science that delivers knowledge about how to program the computers, so as they were able to translate between human languages, for example, Danish and Bulgarian. It may be amazing, but the field of machine translation is almost as old as the invention of the computer itself (Blekhman and Pevzner 2000). In 1949 Warren Weaver, who was a crystallographer, sent a memorandum to The Rockefeller Foundation (an American institution supporting scientific research), in which he demanded that research be started on the automation of translation between natural languages (Arnold et al 1994). Warren Weaver was inspired by cryptographic techniques, which were developed very strongly during the years of The Second World War, and he thought that there existed some fundamental similarities between these cryptographic techniques and the process of translation between human languages (Waibel 2000).

This author does not know if Warren Weaver had a good command of any foreign language, but he claims that the level of Weaver's general linguistic knowledge was rather low. Indeed, it soon appeared that the problem of machine translation is far more complicated and far more harder than Weaver had ever imagined.

Now, after more than fifty years of scientific research in the field of machine translation, a fully-automatic high-quality machine translation system operating on unrestricted text still remains (with some exceptions) an unattainable goal.

So, why is machine translation so difficult, far more difficult than speech recognition or optical character recognition? Is fully-automatic high-quality machine translation for unrestricted text ever possible?

Translation as a Highly Creative Process

To answer the first of the two above questions, let us consider the differences which we can discover when we compare some of the human languages.

First of all, when we study grammatical systems of any natural languages that are not closely related with each other, we easily can see that there exist many more differences than similarities between them (Zue and Glass 2000). For example, let us compare the systems of personal pronouns of Arabic and Hungarian languages.

Personal pronoun system of Hungarian:

Singular	Plural
1. én	1. mi
2. te	2. ti
3. ő	3. ők

Personal pronoun system of Arabic:

Singular	Double	Plural
1. ana	1. nahnu	1. nahnu
2. (m.) anta	2. antuma	2. (m.) antum
2. (f.) anti		2. (f.) antunna
3. (m.) huua	3. huma	3. (m.) hum
3. (f.) hija		3. (f.) hunna

It's clear that the personal pronouns system of Arabic is much more complicated than that for Hungarian. It is caused by the fact that the Hungarian language has no notion of grammatical gender for words. Also, grammatical number in Hungarian can be only singular or plural, whereas in Arabic it can be singular, plural, or double.

So, one can easily see that translating Hungarian personal pronouns into their Arabic equivalents is a hard task. For example, if we want to translate Hungarian pronoun *ők* (in English *they*) into Arabic we must additionally know how many persons are being referred to by this pronoun *ők*. If exactly two persons are being considered then we will use the Arabic word *huma*. But, if there are more than two persons we must additionally know whether they are men or women. If they are men we will use the Arabic word *hum*, otherwise *hunna*.

How do we know how many persons are involved, and whether they are men or women, when the Hungarian word *ők* says nothing about it? The answer is that we know this from the context of the utterance. A human translator can in most cases very easily extract such contextual information, but the full automation of this process remains pure science-fiction.

Quite big differences between human languages can also be noticed when we study their vocabularies. In fact, the vocabulary of each language is an independent and very compound system. If we want to translate, for example, from Chinese into Croatian it is hard work to find in Croatian the equivalents of Chinese words that preserve their original meanings. A human translator copes with an enormous number of lexical holes, that is, words which have no equivalent in the target language and which as such can be translated only by a (sometimes) long description based on their semantics.

This situation is illustrated in Figure 1 where each rectangle is a symbol of some physically existing object or some abstract entity. The rectangles are numerated from 1 to 6. Further, we have two different natural languages: language A and language B.

We can see that in language A, objects 1 and 2 are described only by one common lexical entity, whereas in language B there exist two different lexical entities, one each for object 1 and object 2.

Further, we can notice that the object 3 has no lexical entity in language A and is therefore a lexical hole, whereas in the language B it has its own lexical item.

Objects 4 and 5 in language A are grouped together in one lexical entity and object 6 is a separate lexical entity, while in language B it is otherwise. We notice that objects 5 and 6 form one lexical entity.

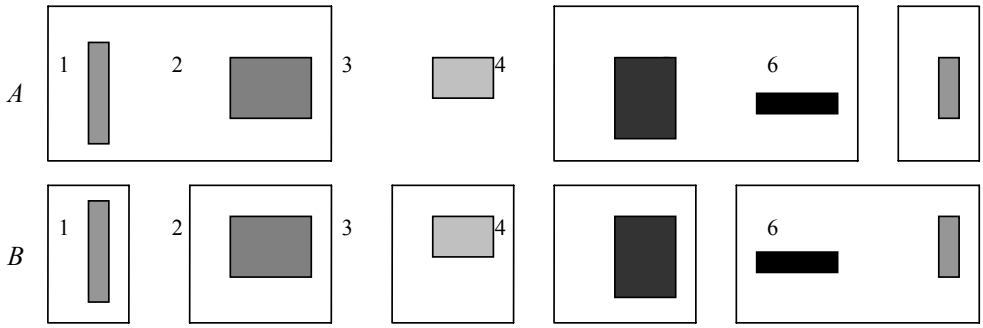

Figure 1. An illustration of the way in which different languages divide reality into lexical items.

A very good example of these possibly rather abstract semantic divisions comes from Swedish language. If we want to translate English word *grandfather* into Swedish, we must additionally know whether this grandfather is a father of a father or a father of a mother. In the first case we should use the Swedish word *farfar* in the other *morfar*, which is illustrated in Figure 2.

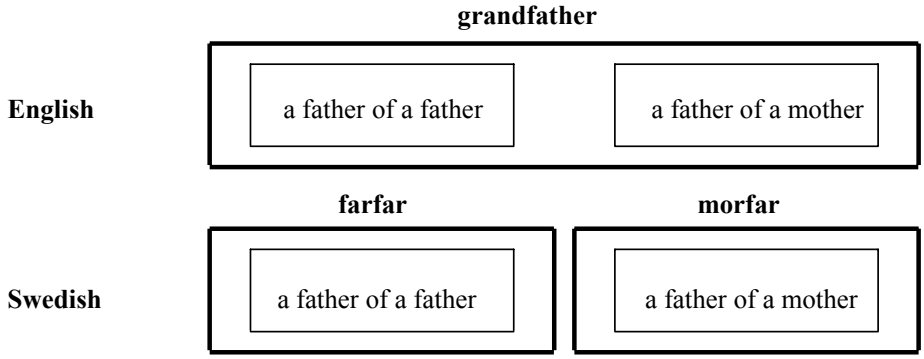

Figure 2. The English word *grandfather* versus Swedish *farfar* and *morfar*.

The most serious problem, which the computer has to cope with in machine translation is the ambiguity of any human language (Baker et al, undated). We can talk of syntactic ambiguity when there exist at least two alternative ways of syntactic analysis of a sentence and of semantic ambiguity when one sentence can be understood in at least two different ways, although the most common is lexical ambiguity. Lexical ambiguity is such a serious problem in the case of machine translation systems because it exists in every natural language and it is really ubiquitous. Indeed, if we open any bilingual dictionary, for example, The Great English-Polish Dictionary, it's very hard to find a word that would only have exactly one meaning. In fact, most of English words have at least two completely different Polish equivalents. So, the question is which one of them the computer should choose while

translating, and where can computer derive the information to establish which is the correct one?

Let us suppose that we have a sentence built from ten different word, and let each of these words has exactly two different meanings. If the computer chooses the equivalents of these words at random this sentence could be translated in 1024 different ways. The probability that acting this way we obtain a correct translation of a whole document built from many such sentences is equal to zero in practice. Moreover, no efficient algorithm that allows for solving this problem is known, and lexical ambiguity can be found in abundance in any human language – below are listed some examples of Polish translation of lexically ambiguous words taken from several languages of the world.

Polish equivalents of the French word **perle** are: 1. perła, 2. paciorek, 3. kapsułka
Polish equivalents of the Spanish word **fondo** are: 1. dno, 2. głębia, 3. tło
Polish equivalents of the Italian word **stufa** are: 1. piec, 2. cieplarnia
Polish equivalents of the German word **Absatz** are: 1. ustęp, 2. obcas, 3. osad, 4. złoże, 5. osadzenie, 6. zbyt
Polish equivalents of the English word **butt** are: 1. beczka, 2. pień, 3. pniak, 4. grubszy koniec, 5. kolba karabinu, 6. płastuga, 7. nasyp za strzelnicą, 8. pośmiewisko, 9. uderzenie głową
Polish equivalents of the Dutch word **boodschap** are: 1. poselstwo, 2. polecenie, 3. wiadomość, 4. zakupy
Polish equivalents of the Swedish word **tomten** are: 1. parcela, 2. plac, 3. krasnoludek
Polish equivalents of the Norwegian word **hytte** are: 1. chata, 2. szałas, 3. buda, 4. huta, 5. kabina
Polish equivalents of the Danish word **løber** are: 1. biegacz, 2. dywanik
Polish equivalents of the Finnish word **kanta** are: 1. podstawa, 2. obcas, 3. stanowisko, 4. baza
Polish equivalents of the Greek word **σκοπός** (*skopos*) are: 1. zamiar, 2. melodia, 3. wartownik
Polish equivalents of the Arabic word **وصل** (*wusal*) are: 1. połączenie, 2. łącze, 3. kontakt, 4. związek, 5. zawias, 6. dodatek

Considering all the above-mentioned factors, translation between natural languages can be seen as a highly creative process. A human translator must have a lot of invention and must know how to deal with situations he had never met before. So, the question is whether it is possible to replace a human being by a computer?

A prominent physicist Roger Penrose in his famous books *New Caesar's Mind* and *The Shadows of the Mind* gave very strong arguments supporting his thesis that the brain operates in a non-algorithmic manner and, because of this fact, a human mind cannot be fully simulated by computer.

So, because we cannot replace a human by a computer does it also mean that a fully-automatic high-quality machine translation for unrestricted text is impossible?

Alan Melby (1999) states that machine translation is headed in the right direction as far as domain-specific approaches using controlled languages are concerned but that further work on fully-automatic high-quality machine translation of unrestricted text is a waste of time and money. To build such systems a real breakthrough in natural language processing (and maybe in the whole filed of information processing) is required. Moreover, such breakthrough will

not be based on any extension of currently known techniques: the electric bulb was not invented because research on the candle had been conducted.

Example-Based Machine Translation

The arguments given by Roger Penrose are very strong and it is no longer possible to ignore them. So, Alan Melby is probably right when he states that it will not be possible to replace a human translator by relying only on currently known techniques. But using these currently known techniques we can still try to come as close as possible to this unattainable goal, which is a fully-automatic high-quality machine translation for unrestricted text. Suppose that during this research we built a machine translation system, which gives a translation of 99% accuracy while operating on an unrestricted text (only 1% of this text need to be edited by a human translator). So can we really say, like Alan Melby, that we had wasted time and money on this research?

Up till now, many totally different approaches to machine translation have been developed, such as: syntactic transfer, interlingua-based machine translation, knowledge-based machine translation, systems based on statistics or neural nets, etc. (Ney et al 2000, Canals et al 2000, Loukachevitch and Dobrov 2000). Of these, example-based machine translation is becoming a serious alternative paradigm, but in most cases it is still an unproven technique in the early research phase (Carbonell et al, undated).

On the other hand, this is not entirely the case. One prominent example comes from Spain. The case of the magazine entitled *Periódico de Catalunya* is interesting because it illustrates probably the first fully operational machine translation system for the translation of unrestricted text that has ever been built, producing nearly 100% satisfactory results while translating from Spanish into Catalan. It is really amazing that this machine translation system is not based on any of the currently known computational linguistics theories. Moreover, it does not analyze the sentence in any way: it only replaces source words (or groups of words) by their target equivalents, just as a spelling-checker would do. The system has a huge dictionary that effectively replaces all linguistic analysis of the source text. The development of the system requires a lot of work: in fact a large team of trained linguists constantly updates the dictionary with new terms, verbs in their different forms and sequences of words of up to six elements. Up till now it has been probably the only practical implementation of a purely unsophisticated machine translation system basing only on a pattern-matching scheme (Rico, undated).

So can this Spanish-Catalan system be an example showing the way to solve the mysterious problem of building a fully-automatic high-quality machine translation system? The answer to this question is not so obvious as one may think. We cannot omit the fact that this Spanish-Catalan system benefits to the great measure from the similarities of the two languages involved in the machine translation process. In fact, the differences between Spanish and Catalan languages are rather minor and are in most cases phonological in nature and only rarely morphological or grammatical.

The results obtained during the development of the Spanish-Catalan machine translation system can be obviously applied to any system which translates between closely related languages. A similarly effective machine translation systems for unrestricted text can probably be built for pairs of languages such as Swedish and Norwegian, Norwegian and Danish, Swedish and Danish, Spanish and Portuguese, German and Dutch, or Finnish and Estonian. But it is doubtful whether a high-quality machine translation system for unrestricted

text can be built in this way that would be able to translate between a pair of totally typologically different and unrelated languages like, for example, Chinese and French.

To imagine how difficult the translation between unrelated languages is, the following experiment was conducted (Majewicz 1989). A sample text written in Polish was taken, the elements (words or phrases) of which were numbered in the following way.

The original Polish text:

Analiza tych dwóch elementów zwyczaju międzynarodowego posiada wielkie znaczenie z uwagi na wyrok Międzynarodowego Trybunału Sprawiedliwości z 29 listopada 1950r. (w sporze między Boliwią a Peru o prawo azylu), który stwierdza, że państwo, które powołuje się na zwyczaj międzynarodowy według art. 38 b musi przeprowadzić dowód, iż powstał on w sposób wiążący drugie państwo.

The order of words in the Polish text:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47

The text was then translated into English, giving the following order of the English equivalents of the elements of the Polish source text.

The English translation of the text:

The analysis of these two elements of the international custom is of a great importance in view of the sentence of the International Court of Justice of the 29th of November 1950 (concerning the dispute between Bolivia and Peru about the right of asylum) which states that a state that is referring to the international custom quoting Article 38b must present the evidence that the custom emerged in a way confining the other state.

The order of words in the English translation:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, **33**, **32**, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47

We can see that in the English translation the word order is almost the same as in the Polish original text. Only two elements (32 and 33) are swapped. This suggests that maybe example-based machine translation technique can be applied successfully to English and Polish. The same text was also translated into Japanese, giving the following word order.

The Japanese translation of the text:

Kokusai kanshu-no kono futatsu-no yoso-no bunseki-wa daisanjuhachi-jo-bi-ni shitagatte kokusai kanshu-o in'yo suru kokka-wa kokusai kan-shu-ga ta-no kokka-o kosoku suru hoho-de sonzai suru to iu shoko-o teishutsu shinakereba naranai to iu koto-o kakunin suru (higoken-ni kansuru boribia peru kan-no funso-ni tsuite-no) senkyuhyakugojunen juichigatsu nijukunichi-no kokusai shiho saibansho-no hanketsu-kara mite hijo-na juyosei-motte iru.

The order of words in the Japanese translation:

6, 5, 2, 3, 4, 1, 36, 35, 37, 34, 33, 32, 30, 28, 43, 46, 47, 45, 44, 42, 40, 39, 38, 26, 24, 23, 22, 19, 21, 21, 18, 17, 16, 15, 12, 14, 13, 11, 10, 8, 9, 7

We can see that the word order in the Japanese translation is totally different from the Polish source text. This does not encourage us to consider using an example-based machine translation technique for such unrelated and typologically different languages. In order to establish whether example based-machine translation is possible between different Indo-European languages, this author conducted some experiments. This author took some samples of the texts belonging to different European languages and translated them manually into Polish using an example-based technique. The chosen translation examples were as short as possible. In fact they could not be any shorter because in this case the Polish translation would be incorrect.

Below are some translation samples, in which the translation-examples have been underlined. The average length of these translation-examples was also calculated (in number of words).

1) Swedish into Polish example-base translation sample:

Jag är på semester på Gotland. (2, 2, 2)
Ja jestem na wakacjach na Gotlandii.
I dag har jag varit och titat på den här stenen. (2, 6, 3)
Dzisiaj byłem popatrzeć na ten oto kamień.
Den är mycket fin med många bilder men det är inga runor på den. (4, 3, 1, 4, 2)
On jest bardzo fajny z wieloma obrazkami ale nie ma żadnych napisów runicznych na nim.
Här har varit fint väder hela tiden. (1, 4, 2)
Tutaj była ładna pogoda przez cały tydzień.
Jag har solat och badad varje dag. (5, 2)
Opalałem i kąpałem się każdego dnia.

The average length of translation examples is 2.81 words.

2) Norwegian into Polish example-base translation sample:

Om kvelden liker vi best å være hjemme og ta det med ro. (2, 3, 3, 1, 4)
Wieczorami wolimy być w domu i mieć spokój.
Vi snakker sammen vi hører på radio eller vi ser på TV. (2, 1, 4, 1, 4)
Jemy razem słuchamy radia albo oglądamy telewizję.
Kanskje drikker far og mor kaffe ved 6-tiden. (1, 5, 2)
Czasami ojciec i matka piją kawę około godziny szóstej.
Vi får melk eller saft. (2, 1, 1, 1)
Dostajemy mleko albo sok.

The average length of translation examples is 2.24 words.

3) Danish into Polish example-base translation sample:

Slottet stammer fra det syttende århundrede. (2, 4)
Zamek ten pochodzi z siódmego wieku.
En berømt forfatter boede og skabte her. (6, 1)

Znany pisarz mieszkał i tworzył tutaj.
Det er udgravninger fra vikingtiden. (3, 2)
To są wykopaliska z czasów Wikingów.

The average length of translation examples is 3.00 words.

4) Dutch into Polish example-based translation sample:

Voor bewoners van Friesland is de moedertall niet Nederlands maar Fries. (4, 5, 2)
Dla mieszkańców Fryzji nie jest językiem ojczystym niderlandzki ale fryzyjski.
Dit is geen dialect het is een cultuurtaal. (4, 4)
To nie jest dialect to jest język kulturalny.
Men kan aan de universiteiten Friese taal- en lettterkunde studeren en men kann daarin doctoraalexamen doen. (2, 3, 5, 1, 2, 3)
Można na uniwersytetach studiować fryzyjską filologię i można zdawać z tej dziedziny egzamin magisterski.
Op de scholen in Friesland bij rechtszittingen in de kerk op vergaderingen van gemeenteraden en van de Provinciale Staten is heel vaak Fries de voertaal. (5, 2, 3, 9, 6)
We fryzyjskich szkołach przy posiedzeniach sądowych w kościele na zgromadzeniach rad okręgowych i rad prowincji bardzo często fryzyjski jest językiem obrad.
Aangezien alle Friezen ook Nederlands kennen is Friesland een tweetalige provincie. (1, 5, 5)
Ponieważ jednak wszyscy Fryzowie znają również język niderlandzki Fryzja jest dwujęzyczną prowincją.

The average length of translation examples is 3.74 words.

5) Spanish into Polish example-based translation sample:

Es un asunto un poco delicado pero te lo voy a contar. (3, 3, 1, 5)
Jest to zagadnienie trochę delikatne ale opowiem ci je.
Hace una semana llegaron los primos de Luis para pasar aquí unos días. (3, 5, 5)
Tydzień temu przyjechali kuzyni Luisa aby pędzić tu kilka dni.
Tú sabes que el estudio de Luis es muy pequeño y por eso les dije que se quedaran en mi casa. (3, 4, 3, 3, 2, 6)
Wiesz że gabinet Luisa jest bardzo mały i dlatego powiedziałem im aby zostali u mnie w domu.
La verdad es que éramos buenos amigos. (4, 3)
Prawda jest taka że byliśmy dobrymi przyjaciółmi.
The average length of translation examples is 4.08 words.

6) Italian into Polish example-based translation sample:

É un fatto abbastanza strano per noi polacchi. (3, 2, 3)
Jest to fakt dosyć dziwny dla nas Polaków.
Da noi le differenze di cui parlo sono meno grandi. (2, 5, 3)
U nas różnice o których mówię są mniejsze.
In Italia invece mentre viaggiavo dal Nord verso il Sud mi sembrava di attraversare non uno ma più Paesi. (3, 2, 5, 8)
Natomiast we Włoszech podczas gdy podróżowałem z północy na południe wydawało mi się że przemierzam nie jeden ale więcej krajów.

Cambiava tutto il paesaggio il clima l'architettura gli abitanti il loro modo di vivere e persino la lingua. (2, 2, 2, 1, 2, 5, 2, 2)
Zmieniało się wszystko pejzaż klimat architektura mieszkańcy ich sposób życia a nawet język.

The average length of translation examples is 3.00 words.

7) French into Polish example-based translation samples:

Parfois le soir après une journée bien remplie j'ai l'impression de n'avoir rien fait. (3, 5, 6)
Czasami wieczorem po całym dniu mocno wypełnionym pracą mam wrażenie że nic nie zrobiłam.
Et je sius fatiguée. (1, 3)
I jestem zmęczona.
Il y a une très grande distance entre le travail que j'ai fait et le résultat. (3, 4, 9)
Jest bardzo duża różnica pomiędzy pracą którą wykonuję a jej rezultatami.
Par example j'ai prepare à manger j'ai fait la vesselle j'ai tout range je me suis occupée de la petite et il me dit tu n'as pas repassé mes chemises. (2, 4, 3, 3, 7, 4, 6)
Na przykład przygotowałam posiłek zmyłam naczynia wszystko poukładałam zajęłam się dzieckiem a on mi mówi nie wyprasowałaś mi koszul.

The average length of translation examples is 4.20 words.

8) German into Polish example-based translation samples:

Mit über 900.000 Einwohner ist Köln heute die drittgrößte Stadt in der Bundesrepublik Deutschland nach Hamburg und München. (4, 6, 4, 4)
Z ponad 900.000 tysiącami mieszkańców jest obecnie Kolonia trzecim co do wielkości miastem w Republice Federalnej Niemiec po Hamburgu i Monachium.
Die Stadt liegt 45 km von der Bundeshauptstadt Bon entfernt. (2, 1, 7)
Miasto leży w odległości 45 km od stolicy związku Bon.
Ihre Geschichte reicht bis in die Römerzeit zurück. (2, 1, 5)
Jego historia sięga aż do czasów rzymskich.
Früher war der Name der Stadt Colonia. (1, 5, 1)
Wcześniej nazwa miasta brzmiała Colonia.

The average length of translation examples is 3.31 words.

9) Slovakian into Polish example-based translation samples:

Jeden z najstarších známych písomných dokladov o škole na Slovensku pochádza z Nitry. (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
Jeden z najstarszych znanych pisemnych dokumentów o szkole na Słowacji pochodzi z Nitry.
Nitra mala pri biskupskej katedrále kapitulnú školu. (1, 1, 1, 1, 1, 2)
Nitra miała przy biskupiej katedrze szkołę zakonną.
Neskôr na tejto škole učili aj vzdelaní laickí učitelia. (1, 3, 1, 1, 1, 2)
Później w szkole tej uczyli również wykształceni nauczyciele świeccy.

The average length of translation examples is 1.16 words.

10) Czech into Polish example-based translation samples:

Česká republika patři mezi nejkrásnější země na světě. (2, 1, 3, 1, 1)
Czeska Republika należy najpiękniejszych krajów na świecie.
Značnou část jejího území tvoří lesy. (1, 1, 1, 1,2)
Znaczną część jej terytorium tworzą lasy.
Jsou zde také rozsáhlé nížiny s loukami pastvinami a poli. (1,1,1,1,1,1,1,1,1,1)
Są tam także obszerne niziny z łąkami pastwiskami i polami.

The average length of translation examples is 1.20 words.

11) Serbo-Croatian into Polish example-based translation samples:

U drugoj sobi stoji krevet moga sina stolić sa dve male stolice ormar za odelo i igračke. (3, 1, 1, 2, 1, 4, 1, 2, 1, 1)
W drugim pokoju stoi łóżko mojego syna stół z dwoma małymi krzesłami szafa na ubrania i zabawki.
Tu se nalazi naš radio aparat. (1, 2, 1, 2)
Tu znajduje się nasz radioodbiornik.
Na zidu vise slike a na prozoru stoje saksije sa cvećem. (2, 2, 1, 2, 2, 2)
Na ścianie wiszą obrazy a na oknie stoją doniczki z kwiatami.

The average length of translation examples is 1.70 words.

From the above translation examples we can see that example-based machine translation between Polish and other Indo-European languages is possible. We can also observe that the more closely related the languages, the shorter the translation examples are.

Conclusions

Based on the example of the Spanish-Catalan machine translation system, which is able to translate an unrestricted text of *Periódico de Catalunya*, and basing on the results of translation experiments conducted by this author, we can draw the conclusion that example-based machine translation is headed in the right direction. In fact, fully-automatic high-quality machine translation for unrestricted text is possible, and we can by no measure say that developing such systems is a waste of time and money, because the positive results of these experiments are obvious (Fukutomi 2000, Murphy 2000, Nyberg et al, Mitamura, Mitamura and Nyberg, undated). But there is one condition that must be fulfilled. The languages between which we want to translate must be related, even if this relationship is not as close as, for example, between Dutch and Polish. But it is very doubtful if the example-based machine translation technique can be applied to languages which are typologically different and wholly unrelated. In such case the problem of building a fully-automatic high-quality machine translation system for unrestricted text is still very far from its final solution, and maybe the further work on such systems using currently known techniques is a pure waste of time and money.

References

Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L. and Sadler, L. (1994) *Machine Translation: An Introductory Guide*, NCC Blackwell, London

Baker, K.L., Franz, A.M., Jordan, P.W., Mitamura, T., Nyberg, E. (undated) *Coping with Ambiguity in a Large-Scale Machine Translation System,* Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA

Blekhman, M. and Pevzner, B (2000) 'First Steps of Language Engineering in the USSR: The 50s through 70s, in *Machine Translation Review*, Issue No. 11, December 2000, pp. 5-7

Canals, R., Esteve, A., Garrido, A., et al (2000) 'InterNOSTRUM: A Spanish-Catalan Machine Translation System', in *Machine Translation Review*, No. 11, December 2000, pp. 21-25

Carbonell, J., Mitamura, T., and Nyberg, E. (undated) *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, ...)*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA

Fukutomi, O. (2000) 'Report on Commercial Machine Translation in a Manufacturing Domain', in *Machine Translation Review*, No. 11, December 2000, pp. 16-25

Loukachevitch, N.V. and Dobrov, B.V. (2000) 'Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems', in *Machine Translation Review*, No. 11, December 2000, pp. 10-20

Majewicz, A. F. (1989) *The languages of the world and their classifying*, Warsaw, Poland

Melby, A. (1999) 'Machine Translation and Philosophy of Language', in *Machine Translation Review*, No. 9, April 1999, pp. 6-17

Murphy, D. (2000) 'Keeping Translation Technology under Control', in *Machine Translation Review*, No. 11, December 2000, pp. 7-10

Ney, H., Nießen, S. Och, F.J., Sawaf, H., Tillmann, C., and Vogel, S (2000) 'Algorithms for Statistical Translation of Spoken Language', in *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1, January 2000, pp. 24-36

Nyberg, E., Mitamura, T., Carbonell, J. *The KANT Machine Translation System: From R&D to Initial Deployment*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA

Rico, C. (undated) *From Novelty to Ubiquity: Computers and Translation at the Close of Industrial Age*, http://www.accurapid.com/journal/15mt2.htm

T. Mitamura (undated) *Controlled Languages for Multilingual Machine Translation*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA

T. Mitamura, Nyberg, E. and Carbonell, J. (undated) *An Efficient Interlingua Translation System for Multi-Lingual Document Production,* Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA

Waibel, A., Geutner, P., Tomokiyo, et al (2000) 'Multilinguality in Speech and Spoken Language Systems', in *Proceedings of the IEEE*, Vol. 88, No. 8, August 2000, pp. 1297-1313

Zue, V.W. and Glass, J.R. (2000) 'Conversational Interfaces: Advances and Challenges', in *Proceedings of the IEEE*, Vol. 88, No. 8, August 2000, pp. 1166-1180

PC-based Machine Translation: an Illustration of Capabilities in Response to Submitted Test Sentences

by

Derek Lewis

The following is an outline of a talk given to the Natural Language Translation Specialist Subgroup of the British Computer Society on 11 December 2001. Members were invited to submit test sentences for translation into a number of languages in order to provide the basis for an informal assessment of the capabilities of PC-based Machine Translation (MT). Two well known and widely available operational packages were used: SYSTRAN PROFESSIONAL (version 3.3, purchased 2000) and GLOBALINK POWER TRANSLATOR PRO (version 6.4, 1998).

SYSTRAN PROFESSIONAL requires 10 MB storage space, plus 16 MB for each installed language pair. The version used was capable of translating between French, German, Italian, Portuguese, Spanish and English, although the demonstration focused on English to German and also provided translations into French. A PRO PREMIUM version of SYSTRAN is also available for English into Japanese and Korean and from Chinese, Japanese, Korean and Russian into English. SYSTRAN operates with three main processing modules: the source language (SL) analysis module is described as able to remember 'uncertainties' established in the initial parsing phase which can be referred to for later analysis. The transfer module is specific to a language-pair, while the target language (TL) synthesis component assigns case, gender and number. There are dictionaries for stem forms (including terms and base forms) and expressions (defined as phrases and so-called 'conditional expressions').

The GLOBALINK POWER TRANSLATOR PRO needs at least 8 MB of RAM (16 MB is recommended), 28-35 MB storage space per language pair and 82 MB for all five language pairs (these are English to and from Spanish, French, German, Italian and Portuguese, and French to German and vice versa).

Below are 117 English sentences which were submitted for translation and a further 21 which a member proposed for translation (into French) because they illustrated cases of syntactic homonymy. The seminar focused on translation into German since this is the language with which I am most familiar and for which I am best able to assess the output. Translations into French, however, are also provided for interest. In the following list S denotes the Systran translation and G the output by Globalink. Reasonable effort was made to update the dictionaries of each system in order to provide the best possible translation (occasionally, updating one item led to a degrading of another, see sentence 102). It must be stressed that the object of the seminar was not to provide a definitive evaluation of the comparative performance of the two systems: there is a substantial body of literature indicating the issues involved and outlining the various parameters which might inform such an exercise in, say, an operational or developmental environment. The results presented here apply only to the sentences submitted, and whether any further conclusions may be drawn on the basis of the output is beyond the scope of this paper; under no circumstances should the sentences be taken to constitute a valid test suite for evaluation purposes. Some sentences exhibited a level of internal syntactic complexity which is known to be beyond the capabilities of MT systems, with predictable results. Others (103, 104, 105, 106, 110) were

evidently provided to see how MT could handle syntactic ambiguity: in the event it appeared that both systems coped reasonably well with the ambiguity, either resolving it satisfactorily (103, 105, 109) or transferring it to the TL (110); however, it is difficult to arrive at a conclusion about a particular issue of ambiguity if the input sentence introduces too many other types of complexity (104, 106). Despite all reservations about judging systems on the basis of a limited number of subjectively chosen or constructed sentences, some comparative assessment of overall translation quality is required. The approach taken was a relative one: for each pair of target language translations I judged which was the better output in all cases where such a judgment could reasonably be made; if there was no evident difference, the translation quality was held to be equal (of course this approach begs entirely the question of whether translation quality was good or bad in absolute terms and therefore avoids making general claims about MT in general). On this basis 30 of the 117 sentences (25.6%) were translated better than Systran, while 24 (20.5%) translated by Systran were judged better than the Globalink output. 63 sentences (53.8%) were held to be of equal quality. In some cases there was a marked difference in quality: in 9 sentences (7.7%) Globalink did much better than Systran, which in turn significantly outperformed Globalink in 4 cases (3.4%). All we can say, therefore, is that Globalink exhibited a slight edge over Systran for this very small corpus of sentences and for this particular language pair and direction. Some sentences were translated very well indeed and a small proportion rendered, by any standards, very badly indeed (84, 90).

On the whole Globalink's dictionaries give the user greater flexibility for updating lexical entries. For example, Systran permits only noun-verb homographs, so *well* cannot be entered as both a noun and as an adverb/adjective. On the other hand, it is easier to enter new items into Systran's dictionaries, although the system's internal lexicons cannot be browsed. The Power Translator includes a Rule Editor for adding phrases and their translations. This tool is based on the user entering items into syntactic frames or rule templates specifying source-to-target transformations of syntactic categories. Here is an example of a rule template for a verb frame for German:

```
// VERB + PARTICLE + DIRECT OBJECT ==> VERB + DIRECT OBJECT
//
// Example: make s.th. up ==> etw. erfinden
//
// SOURCE.1 = make; SOURCE.2 = up; TARGET.1 =erfinden
// ********************************************************************

// RULESET1
// Procedure = Verb Frame; Stage = Frame; Key = SOURCE.1

'SOURCE.1' Particle('SOURCE.2') Obj(SX_Direct)       // invent
   ==> 'TARGET.1';
```

Lines (or parts of lines) beginning with // are comments and are ignored by the program. For each rule there is an example (here: 'make s.th. up ==> etw. erfinden') which illustrates the type of construction to which the rule applies. The user enters the source and target words in the ruleset. Thus inserting *make* for SOURCE.1, *up* for SOURCE.2 and *erfinden* for TARGET.1 should, during the transfer phase, ensure that *make* (which takes a direct object in this sense of 'invent') is translated by *erfinden* and that the particle *up* is discarded. Thus the final rule, without comments, will look as follows:

'make' Particle('up') Obj(SX_Direct ==> 'erfinden';

Experience with rules suggests that they do affect the quality of translation, although the user/dictionary editor cannot assume that the are applied in all cases. Sentence 116 shows that they fail to assist in disambiguating *bus stops*, although this is precisely the example given in the user manual, entered for this demonstration and tested. Consider also the effect of entering *find fault* (English) as *Defekt finden* (German) on translating the following short sentences:

English	**German translation**
a. We find fault.	Wir finden Defekt.
b. We found fault.	Wir fanden Defekt.
c. We find fault with you.	Wir finden, verwerfen Sie mit Ihnen.
d. We found fault with you.	Wir kritisierten Sie.

The entered rule is applied in a and b but not in c. In sentence d the system appears to apply an internal rule which actually delivers a good translation.

For both MT systems idiomatic verb phrases (such as *find fault*, *give up, take off*) can be difficult or impossible to enter in the dictionary (see 67, 71, 75, 99). PP-attachment, an old friend of MT developers, produces predictable results (see 59). To appreciate some of the issues related to PP-attachment consider the following sentences, where the translation of the preposition *on* depends on its attachment to *read* or *book*:

e. I read the book on the beach.
f. I read the book on the church.
g. I read the book on the painting.

Entering a new verb in the lexicon does not necessarily generate the corresponding gerund form (112; see also 81, 83, 84, 85 and 90). Systems can also stumble over the internal structure of NPs (108). For problems of ambiguity, in particular where they lead to bad parses, see 35, 42, 48, 50, 80, 93, 84, 88, 107 and 113-117.

This small corpus provides good examples of the problem of homonymy/polysemy. A lexical form with two or more distinct meanings is said to be homonymous (examples of written homonyms are *bank, train, row*). A polysemous form tends to have two or more related senses (e.g. *foot* of person, page or mountain). Such variation is a common source both of error in translation and of differences in output across systems. For instance, English *table* can correspond to German *Tisch* (furniture) or *Tabelle* (list of data). The list of potential translation homonyms can be extensive: for example, the lexical entry for English *post* in the Power Translator lists at least ten different German translations for the noun and seven for the verb. Without a mechanism of semantic disambiguation it is a matter of chance which item happens to be ranked at the top of the list of translation choices in the dictionaries. The user can select a default translation, depending on the subject domain. Examples of lexical variation include sentences 27, 33, 37, 66, 78, 79, 93, 109.

Other areas of error in the target language output include the position of the negative (36, 40), the translation of the quantifier *some* (58), the dummy pronoun *it* (73), the adverb yet (60), clock times (65), the choice of preposition (81), the distinction between the active and static passive (55), cases after verbs (74), and more complex syntactic transformations (as in *They*

were not permitted, 54). In contrast to English, German readily forms compound NPs as single words but with a complex internal structure. As a result *spare wheel* (49) can be translated accurately by the appropriate term, *Ersatzrad*, if it happens to be present in the lexicon as a semantic unit, or rather less successfully if parsed as an normal adjective + noun construction (*übriges Rad*).

1. Please watch me.
S: Passen Sie mich bitte auf .
G: Beobachten Sie mich bitte.
S: Veuillez m'observer
G: S'il vous plaît regardez-moi.

2. Please pass the list to the Secretary
S: Führen Sie der Sekretärin bitte die Liste.
G: Reichen Sie die Liste bitte zur Sekretärin herüber.
S: Veuillez passer la liste au secrétaire
G: S'il vous plaît laissez-passer la liste au Secrétaire

3. Please don't go there.
S: Bitte gehen nicht. [ignores *there*]
G: Bitte gehen Sie dort nicht.
S: Veuillez ne pas disparaître là.
G: S'il vous plaît n'allez pas là.

4. Is your manager at work?
S: Ist Ihr Manager an der Arbeit?
G: Ist Ihr Manager bei Arbeit?
S: Votre directeur est-il au travail?
G: Est-ce que votre directeur est à travail?

5. He has a bad cold.
[verb phrases such as *have a cold* cannot be entered in SYSTRAN or PT]
S: Er hat eine schlechte Kälte.
G: Er hat eine schlechte Kälte.
S: Il a un mauvais rhume.
G: Il a un mauvais rhume.

6. I'm sorry.
S: Ich bin traurig.
G: Es tut mir leid.
S: Je suis désolé.
G: Je suis désolé.

7. I think the chief lawyer is a colleague of yours.
S: Ich denke, daß der Hauptrechtsanwalt ein Kollege von *Ihrem* ist.
S: Ich denke, daß der Hauptrechtsanwalt ein Kollege von Ihnen ist.
G: Ich glaube, daß der *Haupt Anwalt* ein Kollege von *Ihnen* ist.
G: Ich glaube, daß der Haupt Anwalt *von Ihnen ein Kollege von* ist.
[After entering *of yours* in dictionary]

S: Je pense que l'avocat en chef est un collègue à vous.
G: Je pense que l'avocat principal est collègue du vôtre.

8. I wonder whether you know the answer.
S: Ich wundere, ob Sie die Antwort kennen.
[*know* (intranstive) = *wissen*; *know* + *that* = *wissen*; *know* = direct obj = *kennen*, according to dictionary]
G: Ich frage mich, ob Sie die Antwort wissen. [to change, select from ranked list in dictionary]
S: Je me demande si vous savez la réponse.
G: Je me demande si vous savez la réponse.

9. We want some more airmail envelopes.
S: Wir wünschen mehr Luftpostumschläge.
G: Wir wollen einige mehr Luftpost-Umschläge.
S: Nous voulons encore plus d'enveloppes de par avion.
G: Nous voulons quelques-uns plus d'enveloppes du poste aérienne.

10. When do you require them?
S: Wann benötigen Sie sie?
G: Wann erfordern Sie sie?
S: Quand avez-vous besoin d'elles?
G: Quand est-ce que vous les exigez?

11. The new components do not conform to our requirements..
S: Die neuen Bestandteile passen nicht sich an unsere Anforderungen an.
G: Die neuen Bestandteile passen sich nicht zu unseren Anforderungen an..
S: Les nouveaux composants ne répondent pas à nos exigences.
G: Les nouveaux composants ne conforment pas à nos exigences..

12. Her mother's old coat does not suit her.
S: Alter Mantel ihr Mutter entspricht ihr nicht.
S: Der alte Mantel ihrer Mutter *entspricht* ihr nicht. [After pre-editing: *the old coat of her mother*]
G: Der alte Mantel ihrer Mutter paßt ihr nicht.
S: Le vieux manteau de sa mère ne lui convient pas
G: Le vieux manteau de sa mère ne lui convient pas.

13. That colo(u)r doesn't go with her jacket.
S: Diese Farbe gehört nicht zu ihrer Jacke.
G: Diese Farbe geht nicht mit ihrer Jacke.
S: Ce couleur n'est pas assorti à sa veste.
G: Ce coleur ne va pas avec sa veste.

14. Her new dress matches her coat.
S: Ihr neues Kleid *bringt* ihren Mantel *zusammen*.
S: Ihr neues Kleid paßt ihren Mantel. [After entering *match* = *passen*]
G: Ihre neuen Kleid-*Wettkämpfe* ihr Mantel.
S: Sa nouvelle robe assortit son manteau.

G: Ses nouveaux égaux de la robe son manteau.

15. His grandfather owns a good business.
S: Sein Großvater besitzt ein gutes Geschäft.
G: Seinem Großvater gehört ein gutes Unternehmen.
S: Son grand-père possède de bonnes affaires.
G: Son grand-père possède une bonne affaire.

16. It contains two globes and an atlas.
S: Es enthält zwei Kugeln und einen Atlas.
G: Es enthält zwei Globusse und einen Atlas.
S: Elles contiennent deux globes et un atlas.
G: Il contient deux globes et un atlas.

17. Do you know how to pronounce that word correctly?
S: Können Sie dieses Wort richtig aussprechen?
G: Wissen Sie, wie dieses Wort korrekt auszusprechen ist?
S: Savez-vous prononcer ce mot correctement?
G: Est-ce que vous savez comment prononcer ce mot correctement?

18. They like going round the market.
S: Sie mögen ringsum den Markt gehen.
G: Es macht ihnen Spaß, um den Markt zu gehen.
S: Ils aiment aller autour du marché.
G: Ils aiment le départ autour le marché.

19. We dislike working in the kitchen.
S: Wir lehnen das Arbeiten in der Küche ab.
G: Wir mögen nicht das Arbeiten in der Küche.
S: Nous détestons travailler dans la cuisine.
G: Nous détestons travailler dans la cuisine.

20. Everyone appreciates running water in hot weather.
S: Jeder schätzt laufendes Wasser bei heißem Wetter.
G: Jeder schätzt das Starten von Wasser in heißem Wetter.
G: Jeder schätzt laufendes Wasser in heißem Wetter. [After dictionary update]
S: Chacun apprécie l'eau courante par temps chaud.
G: Tout le monde apprécie eau courante dans temps chaud.

21. How do people know that?
S: Wie wissen Leute die? [*die* should be *das*]
G: Wie wissen Leute das?
S: Comment les gens savent-ils cela?
G: Comment est-ce que les gens savent cela?

21a. How do people know this?
S: Wie wissen Leute dieses.
G:Wie wissen Leute dieses.

S: ??
G:

22. All rivers flow to the sea.
S: Alle Flüsse fließen zum Meer
G: Alle Flüsse fließen zum Meer.
S: Tous les fleuves coulent dans la mer.
G: Toutes les rivières coulent à la mer.

23. I often have extra training on Fridays.
S: Ich habe häufig Extratraining an Freitag.
G: Ich habe extra Schulung freitags oft.
S: J'ai souvent la formation supplémentaire le vendredi.
G: J'ai souvent la formation supplémentaire les vendredis.

24. They do not usually eat cheese.
S: Sie nicht normalerweise essen Käse. [Wrong word order]
G: Sie essen keinen Käse normalerweise.
S: Ils ne mangent pas habituellement du fromage.
G: Ils ne mangent pas de fromage habituellement.

25. Do you ever go to the library?
S: Gehen Sie überhaupt zur Bibliothek?
G: Gehen Sie je zur Bibliothek?
S: Allez-vous jamais à la bibliothèque?
G: Est-ce que vous allez jamais à la bibliothèque?

26. Where does he usually go in the holidays?
S: Wohin geht er normalerweise in die Feiertage?
G: Wo geht er normalerweise in die Feiertage?
S: Où entre-t-il habituellement en vacances?
G: Où est-ce qu'il entre dans les fêtes habituellement?

27. When do you have your next examination?
S: Wann haben Sie Ihre folgende Prüfung?
G: Wann haben Sie Ihre nächste Prüfung?
S: Quand avez-vous votre prochain examen?
G: Quand est-ce que vous avez votre prochain examen?

28. The President addresses the Council tomorrow.
S: Der Präsident spricht zu dem Rat morgen.
G: Der Präsident adressiert morgen den Rat.
S: Le président s'adresse au Conseil demain.
G: Le Président adresse le Conseil demain.

29. I hope you don’t miss the train.
S: Ich hoffe, daß Sie nicht den Zug vermissen.
G: Ich hoffe, daß Sie den Zug nicht verpassen.

S: J'espère que vous ne manquez pas le train.
G: J'espère que vous ne manquiez pas le train.

30. I'll tell you later.
S: Ich erkläre Ihnen später.
G: Ich werde *Sie* später erzählen.
S: Je vous dirai plus tard.
G: Je vous dirai plus tard.

31. This will be the best place *to look for it*.
S: Dieses ist der beste Platz *zum Suchen nach es*.
G: Dies wird die beste Stelle sein, es zu suchen.
S: Ce sera le meilleur endroit pour le rechercher
G: Ce sera la meilleure place pour le chercher.

32. I must calculate the result.
S: Ich muß das Resultat errechnen.
G: Ich muß das Ergebnis kalkulieren.
S: Je dois calculer le résultat.
G: Je dois calculer le résultat.

33. I wish they would not shout so loudly.
S: Ich wünsche, daß sie nicht so laut schreien würden.
G: Ich wünsche, daß sie nicht so laut rufen würden.
S: Je souhaite qu'ils ne crient pas tellement fort.
G: Je souhaite qu'ils ne crient pas si haut.

34. I think we should welcome this suggestion.
S: Ich denke, daß wir diesen Vorschlag begrüßen sollten.
G: Ich glaube, daß wir diesen Vorschlag begrüßen sollten.
S: Je pense que nous devrions faire bon accueil à cette suggestion.
G: Je pense que nous devrions accueillir cette suggestion.

35. I saw her friend *turn* round the corner.
S: Ich sah ihre Freund*umdrehung* ringsum die Ecke.
G: Ich sah ihren Freund um die Ecke drehen.
S: J'ai vu son tour d'ami autour du coin.
G: J'ai vu son ami tourner arrondissez le coin.

36. I cannot let him test the engine.
S: Ich kann nicht ihn die Maschine prüfen lassen.
G: Ich kann ihn nicht lassen den Motor prüfen.
S: Je ne puis pas le laisser examiner le moteur.
G: Je ne peux pas le laisser tester le moteur.

37. I enclosed it in a letter the day before yesterday.
S: Ich umgab sie in einem *Buchstaben* der Tag vor gestern.
G: Ich schloß es in einem Brief der Tag vor gestern ein.

S: Je l'ai enfermé dans une lettre le jour avant hier.
G: Je l'ai joint dans une lettre avant-hier.

38. He did not support your proposal at the meeting.
S: Er stützte Ihren Antrag nicht bei der Sitzung.
G: Er unterstützte Ihren Vorschlag nicht bei der Versammlung.
S: Il n'a pas appuyé votre proposition lors de la réunion.
G: Il n'a pas supporté votre proposition à la réunion.

39. Did you reach an agreement before the meeting?
S: Erreichten Sie eine Vereinbarung vor der Sitzung?
G: Trafen Sie eine Vereinbarung vor der Versammlung?
S: Avez-vous conclu un accord avant la réunion?
G: Est-ce que vous êtes arrivés à un accord avant la réunion?

40. Didn't they search through the register quite recently?
S: Nicht suchten sie durch das Register ziemlich vor kurzem?
S: Suchten sie nicht …. (Afer pre-editing: *Did they not search* ….]
G: Suchten sie ganz vor kurzem durch das Register nicht?
S: N'ont-ils pas recherché par le registre tout à fait récemment?
G: Est-ce qu'ils n'ont pas cherché à travers le registre tout à fait récemment?

41. We enjoyed the grammar lesson.
S: Wir genossen die Grammatiklektion. Nous avons apprécié la leçon de grammaire.
G: Wir genossen die Grammatik-Lehre.
S: Nous avons apprécié la leçon de grammaire.
G: Nous avons aimé la leçon de la grammaire.

42. As soon as the train left he crossed to the other platform.
S: Sobald der Zug verließ, kreuzte er zur anderen Plattform.
G: Sobald die Zug-*Linke*, die er zur anderen Plattform überquerte.
S: Dès que le train est parti il a croisé à l'autre plateforme.
G: Dès que le train est parti il a traversé à l'autre plate-forme.

43. While the inspector was there they did not talk.
S: Während der Prüfer dort war, sprachen sie nicht.
G: Während der Inspektor dort war, redeten sie nicht.
S: Tandis que l'inspecteur était là ils n'ont pas parlé.
G: Pendant que l'inspecteur était là ils n'ont pas parlé.

44. They watched us while the shop remained open.
S: Sie paßten uns auf, während das Geschäft geöffnet blieb.
G: Sie beobachteten uns, während das Geschäft offen blieb.
S: Ils nous ont observés tandis que le magasin restait ouvert.
G: Ils nous ont regardés pendant que le magasin est resté ouvert.

45. He completed it while we walked round the town.
S: Er führte es durch, während wir ringsum die Stadt gingen.
G: Er vervollständigte es, während wir um die Stadt gingen.

S: Il l'a accompli tandis que nous marchions autour de la ville.
G: Il l'a complété pendant que nous avons marché autour la ville.

46. When he was studying English he worked hard.
S: Als er Englisch studierte, arbeitete er stark.
G: Als er Englisch studierte, arbeitete er schwer.
S: Quand il étudiait l'anglais il a travaillé dur.
G: Quand il étudiait anglais il a travaillé difficilement.

47. How did you pass the time when you were young?
G: Wie verbrachten Sie die Zeit, als Sie jung waren?
S: Wie *führten* Sie die Zeit, als Sie jung waren?
S: Comment avez-vous passé le moment où vous étiez jeune?
G: Comment est-ce que vous êtes passés le temps quand vous étiez jeune?

48. They guessed you spoke Russian.
S: Sie schätzten, daß Sie Russen sprachen.
G: Sie rieten Sie *Speiche* Russisch.
S: Ils ont deviné que vous avez parlé russe.
G: Ils ont deviné vous avez parlé russe.

49. She wishes the car had a spare wheel.
S: Sie wünscht, daß das Auto ein Ersatzrad hatte.
G: Sie wünscht, daß das Auto ein übriges Rad hätte.
S: Elle souhaite que la voiture ait eu une roue disponible.
G: Elle souhaite que la voiture eût une roue de rechange.

50. *Well* printed maps are urgently required.
S: *Brunnen* gedruckte Diagramme werden dringend angefordert.
G: Gut gedruckte Landkarten werden dringend erfordert.
S: Des cartes imprimées par bien sont instamment exigées.
G: Les cartes bien imprimées sont exigées d'urgence.

51. Products manufactured in our country are exported all over the world.
S: Die Produkte, die in unserem Land hergestellt werden, werden auf der ganzen Erde exportiert.
G: Produkte, die in unserem Land hergestellt werden, werden in aller Welt exportiert..
S: Des produits construits en notre pays sont exportés partout dans le monde.
G: Les produits fabriqués dans notre pays sont exportés dans le monde entier.

52. He seems confused by the president's speech.
S: Er scheint durch die Rede des Präsidenten konfus.
G: Er scheint verwirrt von der Rede des P Il semble confus par le discours du président.
räsidenten.
S: Il semble confus par le discours du président.
G: Il paraît confus par la parole du président.

53. The safe was stolen before breakfast.
S: Das Safe wurde vor Frühstück gestohlen.
G: Der Safe wurde vor Frühstück gestohlen.

S: Le coffre-fort a été volé avant petit déjeuner.
G: Le coffre-fort a été volé avant petit déjeuner.

54. They *were not permitted* to complete the survey.
S: Sie *wurden nicht die Erlaubnis gehabt*, um die Übersicht durchzuführen.
G: *Sie wurden nicht erlaubt*, die Umfrage zu vervollständigen.
S: Ils n'ont pas été autorisés pour accomplir l'aperçu.
G: Ils n'ont pas été autorisés à compléter l'étude.

55. Not enough maize is grown in this region.
S: Nicht genügend Mais wird in dieser Region angebaut.
G: Nicht wird genug Mais in diesem Gebiet angebaut.
S: Pas assez de maïs est cultivé dans cette région.
G: Pas assez de maïs est grandi dans cette région.

56. Our typewriter is not included in the list.
S: Unsere Schreibmaschine wird nicht in der Liste umfaßt.
G: Unsere Schreibmaschine wird nicht in der Liste eingeschlossen.
S: Notre machine à écrire n'est pas incluse dans la liste.
G: Notre machine à écrire n'est pas incluse dans la liste.

57. A recommendation like this had not been received before.
S: Eine Empfehlung so war nicht vorher empfangen worden.
G: Eine Empfehlung wie dieses war nicht vorher bekommen worden.
S: Une recommandation comme ceci n'avait pas été reçue déja.
G: Une recommandation comme ceci n'avait pas été reçue auparavant.

58. Some modifications were tried out on the vehicle.
S: Etwas Änderungen wurden auf dem Träger ausprobiert.
G: Einige Modifikationen wurden auf dem Fahrzeug ausprobiert.
S: Quelques modifications ont été essayées sur le véhicule.
G: Quelques modifications ont été essayées dehors sur le véhicule.

59. We have already posted the reply *to the Minister*.
S: Wir haben bereits die Antwort *auf den Minister* bekanntgegeben.
G: Wir haben die Antwort schon zum Minister abgeschickt. [after updating to include verb]
S: Nous avons déjà signalé la réponse au ministre
G: Nous avons déjà affiché la réponse au Ministre.

60. Have you looked at the calculation yet?
S: Haben Sie die Berechnung schon betrachtet?
G: Haben Sie die Kalkulation noch angeschaut?
S: Avez-vous regardé le calcul encore?
G: Est-ce que vous avez regardé le calcul cependant?

61. I have explored the area several times.
S: Ich habe den Bereich mehrmals erforscht.
G: Ich habe das Gebiet mehrmals erforscht.

S: J'ai exploré le secteur plusieurs fois.
G: J'ai exploré la région plusieurs fois.

62. I have bought a new car.
S: Ich habe ein neues Auto gekauft
G: Ich habe ein neues Auto gekauft.
S: J'ai acheté une nouvelle voiture.
G: J'ai acheté une nouvelle voiture.

63. He has lived here for several years.
S: Er hat hier für einige Jahre gelebt.
G: Er hat hier mehrere Jahre gelebt.
S: Il a vécu ici pendant plusieurs années.
G: Il a vécu ici pour plusieurs années.

64. We will come back as soon as you have discovered the answer.
S: Wir kommen zurück, sobald Sie die Antwort entdeckt haben.
G: Wir werden zurückkommen, sobald Sie die Antwort entdeckt haben.
S: Nous reviendrons dès que vous découvrirez la réponse.
G: Nous reviendrons dès que vous avez découvert la réponse.

65. By *half past six* the train had already gone.
S: Um *eine Hälfte hinter sechs* war der Zug bereits gegangen.
G: *Durch halbe Vergangenheit sechs, die* der Zug schon gegangen war.
S: Par moitié après six le train était déjà allé.
G: Par à moitié passé six que le train était déjà allé.

66. The ship nearly struck the rock before the storm broke.
S: Das Schiff schlug fast den Felsen an, bevor der Sturm brach.
G: Das Schiff schlug den Stein beinahe, bevor der Sturm brach.
S: Le bateau a presque heurté la roche avant que l'orage se soit cassé.
G: Le bateau presque frappé le roc avant la tempête a cassé.

67. She hoped she had *made* a good impression.
S: Sie hoffte, daß sie einen guten Eindruck *gebildet* hatte.
G: Sie hoffte, daß sie einen guten Eindruck gemacht hatte.
S: Elle a espéré qu'elle avait fait une bonne impression.
G: Elle a espéré qu'elle eût fait une bonne impression.

68. They agreed to abandon the *attempt*.
S: SieWAREN damit einverstanden, den Versuch zu *verlassen*.
Sie stimmten der den Versuch aufgeben zu [After S dictionary update]
G: Sie stimmten überein, den Versuch zu verlassen.
S: Ils ont accepté d'abandonner la tentative.
G: Ils ont consenti à abandonner la tentative.

69. I shouted in order to make him come back again.
S: Ich schrie, um ihn wieder zurückkommen zu lassen.
G: Ich rief, um *ihn zu machen, kommen Sie wieder zurück*.

S: J'ai crié afin de l'inciter à revenir encore.
G: J'ai crié pour le faire revenez encore.

70. The radio began to make a strange noise.
S: Der Radio fing an, merkwürdige Geräusche zu bilden.
G: Das Radio fing an, ein seltsames Geräusch zu machen.
S: La radio a commencé à faire un bruit étrange.
G: La radio a commencé à faire un bruit étrange.

71. He never came to *like* music.
S: Er kam nie zur *wie*musik
G: Er kam nie, um Musik zu mögen.
S: Il n'est jamais venu en musique pareille.
G: Il n'est jamais venu aimer musique.

72. He was not pleased to see her there.
S: Er freute sich nicht, sie dort zu sehen.
G: *Ihm wurde nicht gefallen, um* sie dort zu sehen.
S: Il n'était pas heureux de la voir là.
G: Il n'a pas été heureux de la voir là.

73. *It* happened to be the right answer.
S: *Sie* geschah, die rechte Antwort zu sein.
G: Es passierte, um die richtige Antwort zu sein.
S: Elle s'est avérée justement être la bonne réponse.
G: Il s'est arrivé être la bonne réponse.

74. Her brother has a lot of news to tell you.
S: Ihr Bruder hat eine Menge Nachrichten, Ihnen zu erklären
G: Ihr Bruder hat viele Nachrichten, um *Sie* zu erzählen.
S: Son frère a beaucoup de nouvelles de vous dire
G: Son frère a beaucoup de nouvelles pour vous dire.

75. They want me to tell her to do it.
S: Sie wünschen *mich bitten sie*, es zu tun
G: Sie wollen, daß ich ihr auftrage, es zu machen.
S: Ils veulent que je lui dise de le faire
G: Ils veulent que je lui dise de le faire.

76. My cousin seems anxious to share it with us.
S: Mein Vetter scheint besorgt, es mit uns zu teilen
G: Mein Cousin scheint besorgt, es mit uns zu teilen.
S: Mon cousin semble impatient de les partager avec nous
G: Mon cousin paraît inquiet de le partager avec nous.

77. He is too impatient to learn it properly.
S: Er ist zu ungeduldig, es richtig zu erlernen.
G: Er ist zu ungeduldig, *darum* richtig zu lernen.

S: Il est trop impatient pour l'apprendre correctement.
G: Il est trop impatient de l'apprendre correctement.

78. Why are they moving that table?
S: Warum verschieben sie diese *Tabelle*?
G: Warum bewegen sie diesen Tisch?
S: Pourquoi déplacent-ils cette table?
G: Pourquoi est-ce qu'ils déplacent cette table?

79. They are performing the experiment for a week.
S: Sie führen das Experiment für eine Woche durch.
G: Sie führen das Experiment für eine Woche *auf*.
S: Ils exécutent l'expérience pendant une semaine.
G: Ils exécutent l'expérience pour une semaine.

80. Isn't she getting married?
S: Wird nicht sie *Erhalten verbunden*?
Nicht ist sie heiraten? [after S dictionary update: *get married = heiraten*]
G: Wird sie nicht verheiratet?
S: Obtenir n'est-elle pas mariée?
G: Est-ce qu'elle ne se marie pas?

81. I was checking the translation when someone knocked at the door.
S: Ich überprüfte die Übersetzung, als jemand an der Tür klopfte.
G: Ich überprüfte die Übersetzung, als jemand *bei* der Tür klopfte.
S: Je vérifiais la traduction quand quelqu'un a frappé à la porte.
G: Je vérifiais la traduction quand quelqu'un a frappé à la porte.

82. They were always trying to find fault with one another.
S: Sie versuchten immer, Störung miteinander zu finden.
G: Sie versuchten immer, zu finden, *verwerfen Sie mit einander.*
S: Ils essayaient toujours de trouver le défaut entre eux.
G: Ils essayaient toujours de trouver la faute avec l'un l'autre.

83. Her friend continued complaining about the weather.
S: Ihr Freund *anhaltendes* Beschweren über das Wetter.
G: Ihr Freund setzte das Klagen über das Wetter fort.
S: Son se plaindre continué par ami au sujet du temps.
G: Son ami a continué à se plaindre au sujet du temps.

84. I remember you arriving late.
S: Ich *erinnere an Sie, spät anzukommen.*
G: Ich *erinnere mich an Sie das Ankommen spät.*
S: Je me rappelle vous arriver tard.
G: Je me souviens de vous arriver en retard.

85. We noticed two teachers shouting in the corridor.
S: Wir beachteten zwei Lehrer, im Flur zu schreien.
G: Wir merkten zwei Lehrer-*Geschrei* im Korridor.

S: Nous avons noté deux professeurs crier dans le couloir.
G: Nous avons observé deux professeurs qui crient dans le couloir.

86. Cycling down the avenue I saw the accident happen.
S: *Auslaufen* die Allee sah ich den Unfall zu geschehen.
G: Die Avenue entlang *kreisend* ich sah den Unfall passieren.
S: Cycle en bas de l'avenue j'ai vu l'accident se produire.
G: Faire du vélo en bas l'avenue j'ai vu l'accident se passer.

87. He's afraid of travel(l)ing abroad.
S: Er hat vor auswärts reisen Angst.
G: Er hat Angst vor dem Reisen im Ausland.
S: Il a peur de travel(l)ing à l'étranger.
G: Il a peur de travel(l)ing à l'étranger.

88. You have the opportunity of meeting them tomorrow.
S: Sie haben die Gelegenheit der *Sitzung* sie morgen.
G: Sie haben die Gelegenheit vom *Treffen* von ihnen morgen.
S: Vous avez l'occasion de la réunion ils demain.
G: Vous avez l'occasion de les rencontrer demain.

89. A rattling car makes an irritating noise.
S: Ein ratterndes Auto bildet irritierenden Geräusche.
G: Ein klapperndes Auto macht ein ärgerliches Geräusch.
S: Une voiture de cliquetis fait un bruit irritant.
G: Une crépitant voiture fait un bruit irritant.

90. Walking keeps people fit.
S: *Gehenunterhalt-Leutesitz*
G: *Das Gehen Unterhalt-Leute paßt*.
S: Ajustement de personnes de subsistances de marche
G: La marchant crise des gens des nourritures.

91. If she comes, please give her this message.
S: Wenn sie kommt, bitte geben ihr diese Anzeige.
G: Wenn sie kommt, geben Sie ihr diese Mitteilung bitte.
S: Si elle vient, veuillez lui donnent ce message.
G: Si elle vient, s'il vous plaît donnez-lui ce message.

92. If you intend to stay, we would like to know soon.
S: Wenn Sie beabsichtigen zu bleiben, möchten wir bald wissen.
G: Wenn Sie beabsichtigen, zu bleiben, würden wir bald gern wissen.
S: Si vous avez l'intention de rester, nous voudrions savoir bientôt.
G: Si vous projetez de rester, nous aimerions savoir bientôt.

93. If the generator develops a fault who will look after it?
S: Wenn der Generator eine Störung entwickelt, die um ihn kümmert?
G: Wenn der Generator eine Schuld entwickelt, die sich darum kümmern wird?

S: Si le générateur développe un défaut qui s'occupera de lui?
G: Si le générateur développe une faute qui s'occupera de lui?

94. Even if there is bitter fighting, the world will not end.
S: Selbst wenn es bitteres Kämpfen gibt, *beendet* die Welt nicht
G: Auch wenn es bitteres Streiten gibt, wird die Welt nicht enden.
S: Même s'il y a combat amer, le monde ne finira pas
G: Même s'il y a le combat amer, le monde ne terminera pas.

95. Unless he is cured rapidly, he may die.
S: Es sei denn er schnell kuriert wird, kann er sterben
G: Außer wenn er schnell geheilt wird, stirbt er vielleicht.
S: À moins qu'il soit guéri rapidement, il peut mourir
G: À moins qu'il soit guéri rapidement, il peut mourir.

96. If you believe that, you are making a mistake.
S: Wenn Sie dem glauben, machen Sie einen Fehler.
G: Wenn Sie glauben, *daß*, Sie machen einen Fehler.
S: Si vous croyez cela, vous faites une erreur.
G: Si vous croyez que, vous faites une erreur.

97. Every time one valve opens, the other valve closes.
S: Jedesmal wenn ein Ventil sich öffnet, schließt das andere Ventil.
G: Jedes Mal wenn ein Ventil öffnet, schließt das andere Ventil.
S: Chaque fois qu'une valve s'ouvre, l'autre valve se ferme.
G: Chaque fois une valve ouvre, les autres fins de la valve.

98. As he has recommended it, it ought to be faultless.
S: Wie er es empfohlen hat, soll es tadellos sein.
G: *Als* er es empfohlen hat, sollte es fehlerfrei sein.
S: Comme il l'a recommandée, elle doit être parfaite.
G: Comme il l'a recommandé, ce devrait être sans défaut.

99. If the jet took off on schedule, they will be there by now.
S: Wenn das Düsenflugzeug auf Zeitplan *sich entfernte*, sind sie dort jetzt.
G: Wenn das Düsenflugzeug auf Zeitplan startete, werden sie inzwischen dort sein.
S: Si le gicleur décollait dans les délais, ils seront là près maintenant
G: Si le jet était parti sur programme, ils seront maintenant là.

100. If she were not so obstinate she might be more successful.
S: Wenn sie nicht also *war,* hartnäckig konnte sie erfolgreicher sein
G: Wenn sie nicht so hartnäckig wäre, könnte sie erfolgreicher sein.
S: Si elle n'étaient pas aussi obstiné elle pourrait être plus réussie.
G: Si elle n'était pas si obstinée elle peut être plus prospère.

101. If he had looked under the mat he would have found the key.
S: Wenn er unter der Matte *geschauen* hatte, würde er den Schlüssel gefunden haben.
G: Wenn er unter der Matte gesehen hätte, hätte er den Schlüssel gefunden.

S: S' il avait regardé sous la natte il aurait trouvé la clef
G: S'il avait regardé sous le tapis il aurait trouvé la clef.

102. Have you goods to declare?
S: Haben Sie die Waren zum Erklären?
S: *Lassen* Sie Waren zum Verzollen? [after dictionary update]
G: Haben Sie Sie Güter, um zu erklären?
G: Haben Sie Sie Waren zum Verzollen? (after dictionary update)
S: Vous avez- des marchandises à déclarer?
G: Ayez-vous marchandises pour déclarer?

103. This taxi was bought by a man with a Spanish accent and packed with high explosives.
S: Dieses Taxi wurde von einem Mann mit einem spanischen Akzent gekauft und verpackt mit hochexplosiven Sprengstoffen.
G: Dieses Taxi wurde von einem Mann mit einem spanischen Akzent gekauft und gerammelt volle mit hohen Sprengstoffen.
S: Ce taxi a été acheté par un homme avec un accent espagnol et emballé avec de hauts explosifs.
G: Ce taxi a été acheté par un homme avec un accent espagnol et plein avec les hauts explosifs.

104. I was a bit worried alligator meat would put people off because they are such ugly creatures, but it's gone down really well.
S: Ich war ein Spitze gesorgtes Krokodil, das Fleisch Leute weg setzen würde, weil sie solche häßliche Geschöpfe sind, aber es wird unten wirklich gut gegangen.
G: Ich war ein Stückchen, das beunruhigt wird, daß Alligator-Fleisch Leute verschieben würde, weil sie solche häßliche Kreaturen sind, aber es ist echt gut hinuntergegangen.
S: J'étais un alligator inquiété par peu que la viande mettrait des personnes au loin parce qu'elles sont de telles créatures laides, mais elle est descendue vraiment bien.
G: J'étais la viande de l'alligator un peu inquiète dégoûterait des gens parce qu'elles sont de telles créatures laides, mais il est vraiment bien descendu.

105. Police arrested the women and children when they broke into a house in Camberley in Surrey.
S: Polizei hielt die Frauen und die Kinder fest, als sie in ein Haus in Camberley in Surrey einbrachen.
G: Polizei verhaftete die Frauen und die Kinder, als sie in ein Haus in Camberley in Surrey brachen.
S: La police a arrêté les femmes et les enfants quand elles ont pénétré par effraction dans une maison dans Camberley dans Surrey.
G: La police a arrêté les femmes et enfants quand ils sont entrés de force dans une maison dans Camberley dans Surrey.

106. Ministers want to outlaw private clubs which refuse to admit women as part of their commitment to promoting equality.
S: Minister möchten private Vereine ächten, die ablehnen, Frauen als Teil ihrer Verpflichtung zur Förderung von Gleichheit zuzulassen.

G: Minister wollen private Klubs ächten, die ablehnen, Frauen als Teil ihres Engagements zum Fördern von Gleichheit zuzugeben.
S: Les ministres veulent proscrire les clubs privés qui refusent d'admettre des femmes en tant qu'élément de leur engagement à favoriser l'égalité.
G: Les ministres veulent bannir des clubs privés qui refusent d'admettre des femmes comme partie de leur engagement encourager l'égalité.

107. Foot heads arms body.
S: Fuß geht Armkörper voran.
G: Fuß *Köpfe*-Arme-Körper.
S: Le pied dirige le corps de bras.
G: Le pied corps des bras des têtes.

108. Silent screen actor.
S: Leiser Schirmschauspieler.
S: Stummfilmschauspieler [after entering *silent screen = Stummfilm*]
G: Schweigsamer Bildschirm-Schauspieler.
S: Acteur de cinéma silencieux.
G: L'acteur de l'écran silencieux.

109. 30,000 dead patients' organs
S: 30.000 Organe der toten Patienten
G: die Organe von 30,000 toten Patienten
S: 30.000 organes des patients morts
G: les organes de 30,000 malades morts

110. A plea for help from the Prime Minister
S: Eine Bitte für Hilfe vom Premierminister
G: Eine Bitte für Hilfe vom Premierminister
S: Une intervention en faveur d'aide du premier ministre
G: Une défense pour aide du premier ministre

111. Women get more firsts at university than men.
S: Frauen erhalten mehr Ersten an der Universität als Männer.
G: Frauen *holen* mehr ersten an *Universität* als Männer.
S: Les femmes atteignent plus de premiers l'université que des hommes
G: Les femmes obtiennent premier à université plus qu'hommes.

112. Arrangements for vetting publications by former spies.
S: Vorbereitungen für vetting Publikationen durch ehemalige Spione.
G: Anordnungen für das Prüfen von Veröffentlichungen durch ehemalige Spione.
S: Arrangements pour des publications de contrôle par d'anciens espions.
G: Arrangements pour examiner des publications par les espions du fondateur.

113. I cannot bear children.
S: Ich kann nicht Kinder tragen.
G: Ich kann keine Kinder gebären.
S: Je ne puis pas soutenir des enfants.

G: Je ne peux pas porter d'enfants.

114. I'll be right back.
S: Ich bin rechte Rückseite.
S: Ich bin gleich zurück. [After dictionary update]
G: Ich werde zurück recht haben.
G: Ich werde gleich züruck sein. [After dictionary update]
S: Je serai bon dos.
G: J'aurai raison en arrière.

115. Census takers wanted to count population.
S: Zählungabnehmer *wollten* Bevölkerung zählen.
G: Volkszählung-Nehmer *wollten* Bevölkerung zählen.
S: Les preneurs de recensement ont voulu compter la population.
G: Les preneurs du recensement voulaient compter la population.

116. The bus stops at the bus stops.
G: Die Bushaltestelle bei den Bushaltestellen.
S: Die Busanschläge an den Busanschlägen.
S: Die Bushaltestellen an den Bushaltestellen. [after entering *bus stop*]
S: Les arrêts d'autobus aux arrêts d'autobus.
G: L'autobus arrête aux arrêts de l'autobus.

117. The bus stopped at the bus stops.
G: Die Bushaltestelle bei den Bushaltestellen.
S: Der Bus gestoppt an den Busanschlägen.
S: Bushaltestelle an den Bushaltestellen [after entering *bus stop*]
S: L'autobus arrêté aux arrêts d'autobus
G: L'autobus a arrêté aux arrêts de l'autobus.

The following sentences were submitted because they illustrate instances of syntactic ambiguity, mainly between verb and noun (e.g. *wound*) but also between verb and adjective (e.g. *intimate*) or even noun and (comparative) adjective (*number*). Both systems perform well and fairly equally inasmuch as in most cases they distinguish the basic syntactic categories underlying the ambiguity (with the notable exception of *number* in sentence 19). In some instances the general category (e.g. verb) may be identified but the inflectional form of the category is wrong (as in *présentez* in sentence 7). In other cases, although the right category is parsed, the lexical item within the category is incorrect (e.g. 1, 7, 12 and probably also 18) – an 'error' which is endemic to MT systems which are not tailored to a domain.

SENTENCES illustrating SYNTACTIC HOMONYMY

1. The bandage was wound around the wound.
German – S: Der Verband wurde um die Wunde verwundet.
German – G: Der Verband wurde um die Wunde gewunden.
French – S: Le bandage a été enroulé autour de la blessure.
French – G: Le pansement a été enroulé la blessure autour.

1. The farm was used to produce produce.
German – S: Der Bauernhof wurde benutzt, um Erzeugnis zu produzieren.
German – G: Der Bauernhof wurde benutzt, um Produkt zu produzieren.
French – S: La ferme a été employée pour produire le produit.
French – G: La ferme a été utilisée pour produire le produits alimentaires.

1. The dump was so full that it had to refuse more refuse.
German – S: Das Dump war so voll, daß es mehr Abfall ablehnen mußte.
German – G: Die Müllkippe war so voll, daß es mehr Müll ablehnen mußte.
French – S: La décharge était si pleine qu'elle ait dû refuser plus d'ordures.
French – G: La décharge était si pleine qu'il a dû refuser plus de déchets.

2. We must polish the Polish furniture.
German – S: Wir müssen die polnischen Möbel polieren.
German – G: Wir müssen die polnischen Möbel polieren.
French – S: Nous devons polir les meubles polonais.
French – G: Nous devons polir le mobilier polonais.

3. He could lead if he would get the lead out.
German – S: Er könnte führen, wenn er die Leitung heraus erhalten würde.
German – G: Er könnte führen, wenn er das Blei herausholen würde.
French – S: Il pourrait mener s' il obtiendrait le fil dehors.
French – G: Il pourrait mener s'il sortirait le rôle principal.

4. The soldier decided to desert his dessert in the desert.
German – S: Der Soldat entschied sich, seinen Nachtisch im Ödland zu verlassen.
German – G: Der Soldat entschied sich, seinen Nachtisch in der Wüste zu verlassen.
French – S: Le soldat a décidé d'abandonner son dessert dans le désert.
French – G: Le soldat a décidé d'abandonner son dessert dans le désert.

5. Since there is no time like the present, he thought it was time to present the present.
German – S: Da es keine Zeit wie das Geschenk gibt, dachte er, daß es Zeit war, das Geschenk
German – G: Seitdem gibt es keine Zeit wie die Gegenwart, er glaubte, daß es Zeit zu war, präsentieren Sie die Gegenwart.
French – S: Puisqu'il n'y a plus de temps comme le présent, il a pensé qu'il était temps de présenter le présent.
French – G: Depuis il n'y a pas de temps comme le présent, il pensait que c'était temps à présentez le présent.

6. A bass was painted on the head of the bass drum.
German – S: Ein Baß wurde auf dem Kopf der Baß-Trommel gemalt.
German – G: Ein Baß wurde auf dem Kopf von der Trommel gemalt.
French – S: Une basse a été peinte sur la tête du tambour bas.
French – G: Une basse a été peinte sur la tête du bas tambour.

7. When shot at, the dove dove into the bushes.
German – S: Als Schuß an, die Taube in die Büsche tauchte.
German – G: Wenn bei geschossen hat, die Taube sprang in die Büsche.

French – S: Quand le projectile à, la colombe a plongé dans les buissons.
French – G: Quand a tiré à, la colombe plonge dans les buissons.

8. I did not object to the object.
German – S: Ich wendete nicht gegen den Gegenstand ein.
German – G: Ich wider setzte mich dem Gegenstand nicht.
French – S: Je ne me suis pas opposé à l'objet.
French – G: Je n'ai pas protesté contre l'objet.

9. The insurance was invalid for the invalid.
German – S: Die Versicherung war für das unzulässige unzulässig.
German – G: Die Versicherung war für den Körperbehinderten ungültig.
French – S: L'assurance était inadmissible pour l'inadmissible.
French – G: L'assurance était invalide pour l'invalide.

10. There was a row among the oarsmen about how to row.
German – S: Es gab eine Reihe unter oarsmen über, wie man rudert.
German – G: Es gab eine Reihe unter den Ruderern, wie zu rudern ist.
French – S: Il y avait une rangée parmi oarsmen au sujet de la façon ramer.
French – G: Il y avait une ligne parmi les rameurs au sujet de comment ramer.

11. They were too close to the door to close it.
German –S: Sie waren auch nah an der Tür zum Schließen es.
German – G: Sie waren auch in der Nähe von der Tür, darum zu schließen.
French – S: Ils étaient trop près de la porte pour le clôturer.
French – G: Ils étaient près de la porte pour le fermer aussi.

12. The buck does funny things when the does are present.
13. German – S: Der Dollar tut lustige Sachen, wenn sind anwesend.
German – G: Der Bock macht lustige Sachen wenn das macht, ist anwesend.
French – S: Le mâle fait des choses drôles quand sont présents.
French – G: Le mâle fait des choses drôles quand le fait est présent.

14. A seamstress and a sewer fell down into a sewer line.
German – S: Eine Näherin und ein Abwasserkanal fielen unten in eine Abwasserkanallinie.
German – G: Eine Näherin und ein Abwasserkanal fielen in eine Abwasserkanal-Linie herunter.
French – S: Un ouvrière couturier et un égout sont tombés vers le bas dans une ligne d'égout.
French – G: Une couturière et un égout sont tombés dans une ligne de l'égout.

15. To help with planting, the farmer taught his sow to sow.
German – S: um beim Errichten zu helfen, unterrichtete der Landwirt seinen Abstichgraben zu säen.
German – G: Mit dem Einpflanzen zu helfen, brachte der Bauer seiner Sau bei, zu säen.
French – S: Pour aider avec la plantation, le fermier a enseigné sa truie à semer.
French – G: Pour aider avec planter, le fermier a appris sa truie pour semer.
16. The wind was too strong to wind the sail.
German – S: Der Wind war zu stark, das Segel zu wickeln.
German –G: Der Wind war zu stark, um das Segel zu winden.

French – S: Le vent était trop fort pour enrouler la voile.
French – G: Le vent était trop fort pour enrouler la voile.

17. After a number of injections my jaw got number.
German – S: Nach einer Anzahl von Einspritzungen erhielt mein Kiefer Zahl.
German – G: Nach einer Anzahl von Einspritzungen bekam mein Kiefer Zahl.
French – S: Après un certain nombre d'injections ma mâchoire a obtenu le nombre.
French – G: Après que plusieurs injections que ma mâchoire a obtenu à nombre.

18. Upon seeing the tear in the painting I shed a tear.
German – S: Nach dem Sehen des Risses im Anstrich verschüttete ich einen Riß.
German – G: Auf dem Sehen der Träne im Gemälde vergoß ich eine Träne.
French – S: En voyant la larme dans la peinture j'ai jeté une larme.
French – G: Sur voir la larme dans le tableau je verse une larme.

19. I had to subject the subject to a series of tests.
German – S: Ich mußte abhängig von einer Reihe Tests unterwerfen.
German – G: Ich mußte einer Folge von Prüfungen das Thema aussetzen.
French – S: J'ai dû soumettre sujet à une série d'essais.
French – G: J'ai dû soumettre le sujet à une série d'épreuves.

20. How can I intimate this to my most intimate friend?
German – S: Wie kann vertrautes I dieses zu meinem vertrautesten Freund?
German – G: Wie kann ich dieses zu meinem vertrautesten Freund andeuten?
French – S: Comment ose-t-il I intime ceci à mon ami plus intime?
French – G: Comment est-ce que je peux intimer ceci à mon ami le plus intime?

Semi-Automatic Construction of Multilingual Lexicons

by

Lynne Cahill
ITRI, University of Brighton
Lewes Road, Brighton BN2 4GJ, UK
Lynne.Cahill@itri.bton.ac.uk

Abstract

The construction of lexicons for NLP applications is a potentially very expensive task, but a crucially important one, especially in multilingual applications. The automation of the task from generic data sources or corpora is as yet largely impractical for most *applied* systems. In the paper we describe a methodology for the semi-automation of the task, used in the CLIME project to develop bilingual lexicons for generation in a restricted domain. We go on to discuss ways in which the same methodology has been used to develop lexicons for a range of applications.

Introduction

Despite a large variety of research in recent years addressing issues of the construction of a large lexical resource in a range of languages, it is still the case that most NLP applications do not make use of such resources, but produce tailor-made lexicons for each application. Projects such as ACQUILEX (Copestake et al, 1995), GENELEX (GENELEX Consortium, 1994), EDR (EDR, 1990) and MULTILEX (MULTILEX (1993)) have made great advances in the creation of lexical resources, but practical applied NLG systems, for example, almost invariably make use of relatively small, manually produced specialised lexicons (Cahill, 1998b). We stress here that we are not addressing lexicon building for the purpose of MT, but for other multilingual NLP tasks, namely NLG and NLU. As we shall we, it is often the case in practical NLP tasks that sophisticated theories of semantic relations are not required for adequate performance, in contrast to MT.

There has been a significant amount of work on the structuring, development and maintenance of lexicons for NLP, particularly in the tradition of non-monotonic inheritance. Daelemans and Gazdar (1992) and Briscoe et al. (1993) bring together much of this work on the application of inheritance networks to lexical description, while Cahill and Evans (1990) discusses the issue in relation to the practical goal of making lexicons more portable and extendable.

Other discussions of the development of lexical resources include work on extraction of information from corpora, such as Garside et al (1997); and work of the extraction of information from machine-readable dictionaries, such as Boguraev and Briscoe (1989). However, what is required for the application we have in mind is a semantically much less complex set of lexical information that nevertheless would benefit from shared cross-linguistic information.

In this paper we discuss the methodology we adopted in developing the lexicons needed for an applied NLG system and the reasons for it. This methodology involved a combination of manual and automated development and has resulted in a set of tools that will enable a non-

linguist domain expert to enter the required lexical information to port the lexicon to a new language. We first look at the particular lexical requirements for the CLIME system interface. We then discuss the approach we adopted in the development of English and French lexicons for the CLIME interface before considering similar approaches to lexicons for different NLP tasks. We argue that this type of approach is the most likely way forward in exploiting the wide range of generic lexical resources in NLP applications, as it permits the system developer to combine any number of distinct resources while also tailoring the output to the particular application at hand.

The CLIME Project requirements

The CLIME project is developing a legal reasoning system which can be used by ship surveyors to query a database of legal regulations. The user interface to this is the WYSIWYM (Power, Scott and Evans, 1998) system, which is implemented primarily in ProFit (an extension of Prolog). The user formulates questions by manipulating on-screen texts. These texts contain spans which can either optionally or obligatorily be expanded by the use of menus. In the domain we are modelling, the maritime domain, there are around 3300 *concepts* that have been identified by our partners at the University of Amsterdam as occurring in the portion of the rules they have modelled to date. Each of these concepts needs a lexical entry, providing the syntactic and realisational information needed to generate sentences about the concept. Given the presence of a concept *bilge pump* in the ontology of the system, the WYSIWYM interface will allow the user to phrase such questions as *What is a bilge pump?*, *What are all the parts of a bilge pump*, *What are the things connected to a bilge pump?* and so on. When the answers to the question are returned by the other modules of the system, the response is generated in the chosen language by a back-end generation module.

We need lexical entries for these concepts in both English and French, but we do not require any subtle semantic information for the range of questions that the user can sensibly ask the system. We simply need one form for English and one for French for each concept.

The task of finding simple one-to-one, domain specific translations of the concept set we wanted to represent proved more difficult that we had hoped. On-line dictionaries could be found which gave us the translations we (thought we) wanted, but only amongst several others which we clearly didn't want. In addition, we found that *we* didn't always know which of the translations returned we wanted - this was knowledge that only experts in the domain could reliably provide.

It must be stressed that the implementation of the system makes certain simplifying assumptions about the differences between English and French that prove acceptable in the current application, but which would not be acceptable in an application to perform a different NLP task, such as Information Extraction. These assumptions result in virtually identical grammars for English and French, grammars that are sufficient to generate the limited range of language required for this interface. With the exception of certain rules for the handling of English plurals, the only differences between the languages are handled in the lexicon, either as word forms or as fixed phrases. It is our assumption that any differences that required more sophisticated grammatical treatment would require a (computational) linguist to implement, while the domain specific lexical forms require a domain expert. However, in the model we propose here, the two tasks are entirely separated, so that porting the lexicon to a different domain, or just extending it, can be performed after the linguist has finished development and the system has been deployed.

The CLIME lexicons

The CLIME system has two NLG modules - one which the user interacts with to compose a query and the second which generates the linguistic version of the answer to the query. As discussed above, the first of these uses the WYSIWYM system (Power, Scott and Evans, 1998), which is implemented in ProFit, an extension of Prolog. The system currently generates English and French, and will shortly be extended to include Italian. The core parts of the lexicons for the NLG modules were entered manually, including the core lexemes for each language - i.e. determiners, common nouns, auxiliaries, fixed phrases for the domain etc. For the core parts to function, however, it is vital that there is a lexical entry for each concept in the domain model. The domain model for the NLF is derived from an ontology (the Legal Knowledge Repository or LKR) that is used by the legal reasoning system. We subsequently devised a system for automatically extending all the lexicons required to cover all of the concepts in the ontology.

The ontology comes to us in HTML format. From this we derive two things: a database consisting of subtype definitions. To this database, we manually added French translation of the concepts (the translations were provided by our project partner, Bureau Veritus in Paris), together with their gender. There was no obvious alternative to this manual translation effort, because the translations we required were very domain specific. As we discussed above, we could not find any machine-readable dictionary that could provide for us the single most appropriate translation for terms such as 'bilge pump' or 'horizontal bulkhead'. This is an area where domain experts are needed, but we did not want to force these experts to get their hands dirty entering the translations into a structured lexicon, nor did we want to have to enter all of the (3000+) translations manually ourselves. Thus, we opted for the best compromise, where the French experts entered the translations into a simple database (in fact it was done in an Excel spreadsheet which we subsequently dumped out into ASCII) from which we could then automatically generate the structured lexicons required.

From this database, a set of inheritance-based hierarchically structured lexicons, were produced, with the top structure manually crafted and the bulk of the lexemes at the leaves automatically generated (these lexicons were defined in the lexical representation language DATR (Evans and Gazdar, 1996)). These included the sharing of cross-linguistic information. In contrast to the PolyLex model (Cahill and Gazdar, 1999), in which shared information is contained in a separate multilingual hierarchy, the default hierarchy in this case was the English one. The main reason for this was simply the practical consideration that we had started with the English lexicon and then extended it to French. However, this also carries the benefit of being able to use the English word where the French translation is not available. Although not an ideal situation, it was felt that it was better to have an English term appearing in the French text than to have the system fail to produce a text at all if some French translations were missing. It is also the case in this particular application that many of the concepts are actually abbreviations (e.g. 'cbt', 'ice_i'), for which it does not make sense to have a translation.

The next stage of generation of the lexicons combines the hierarchically organised information with the Prolog subtype information to construct ProFit entries as required by the WYSIWYM system. The subtype information is used to determine whether a noun is mass or count - subtypes of 'ship', 'equipment', 'system' etc. are count, while subtypes of 'notation', 'state' etc. are mass. (Although this is a simplification, it is one which works a large proportion of the time. All of the automatic lexicon construction described here assumes that

some checking may be necessary to deal with certain lexical exceptions. In some cases there are ways of dealing with this explicitly. For example, the automatic construction of the PolyLex lexicons (Cahill 1998a) produces a separate file for words whose morphological behaviour does not exactly match any of the available classes, while those words are given default morphological values in the automatically produced lexicons.) Let us look at an example lexical entry.

The NLG part of the WYSIWYM system is written in ProFit, and consists of grammar rules that the generator attempts to instantiate by realising the 'right-hand side' where the meaning matches the 'left-hand side'. The lexicon is essentially a set of declarative rules that define sets of feature-value pairs that correspond. In generation terms, this means that we index on (primarily) the meaning feature, and the output is the value of the cset feature. The WYSIWYM lexicon needs entries like the following:

```
word(english,   meaning!cargo_ship &
                syntax!(category!noun &
                        opening!consonant &
                        form!common &
                        noun_type!count) &
                        cset!'cargo ship').
```

Here, the ProFit defines a set of feature/value pairs such as noun_type (feature) and count (value). In the automatically derived section of the lexicon in DATR the corresponding entry for 'cargo ship' in English is:

```
E_Cargo_ship:
    <> == Noun
    <syntax category> == noun
    <opening> == consonant
    <form> == common
    <noun_type> == 'cargo ship'.
```

In French, this is:

```
F_Cargo_ship:
    <> == E_Cargo_ship
    <gender> == masculine
    <cset> == 'navire cargo'
```

From these basic DATR entries, lexical entries are generated for both languages, for two different types of entry that are used for asking different types of question. In addition, the second NLG module requires slightly different lexical entries again, and these too can be generated from the same DATR entries. We therefore generate six separate lexicons from these entries, the main WYSIWYM lexicons, the concept lexicons used by one more specific part of the WYSIWYM interface and the lexicons for the back-end generation.

The whole process is illustrated in figure 1. In the figure, the solid boxes are what we consider to be non-lexical databases or information sources. (Of course, the boundary between these different types of resource are unclear. The HTML LKR, for instance, is not strictly a lexical resource, but it nevertheless contains a large proportion of the information

required by a lexicon.) The dashed boxes are lexicons. The solid arrows between the boxes are fully automatic derivation, while the dashed arrows indicate manual derivation.

OtherNLP applications

The methodology described above can be viewed as having at least two stages: the first moving from a (largely unstructured) database to a more highly structured lexicon and the second from this structured lexicon to an application specific lexicon which may be less structured again, but which may have more highly structured (and programming language specific) individual entries.

In this section we briefly discuss two different lexicon building processes that each undertake one of these two levels. The PolyLex automatic extension process takes the largely unstructured CELEX database to extend the highly structured PolyLex multilingual lexicons. The lexicons for the POETIC project were constructed as highly structured lexicons, from which less structured, application specific lexicons were automatically derived. We shall look at each of these in turn.

The PolyLex lexicons

The aim of the PolyLex project was not to build lexicons for a particular application or application type, but to develop hierarchically structured lexicons that organised the information about related languages in a way that permitted sharing of information across all different levels of linguistic description (Cahill and Gazdar 1999). The resulting lexicons covered morphological, morphophonological and phonological information primarily, with some syntactic and orthographic information. The information common to two or more of the three languages covered - Dutch, English and German - was contained in a *multilingual* hierarchy (this should more properly be described as a set of hierarchies, as the different levels of information tend to be defined in essentially separate, although possibly interacting hierarchies), with the individual language hierarchies inheriting this information by default and overriding it where necessary.

The methodology employed in developing the lexicons was to first develop a core multilingual lexicon including around 300 words for each language. These items were chosen because they were representative of all of the different *morphological* classes, and so they included most of the irregular words of each language. These were developed as default inheritance hierarchies, implemented in the lexical knowledge representation languages, DATR (Evans and Gazdar 1996), with the lexemes as the leaf nodes of the hierarchy. In order to then extend the lexicons to the intended level of 3000 words for each language, it was decided to automatically induce the lexical entries from a combination of the CELEX lexical database and manual translations from English into German and Dutch. This manual translation was chosen again because of the difficulty in finding automatically simple one-to-one translations. The translations were done by bi-lingual speakers of English/Dutch or English/German who could most reliably give the most straightforward translations of the list of common words.

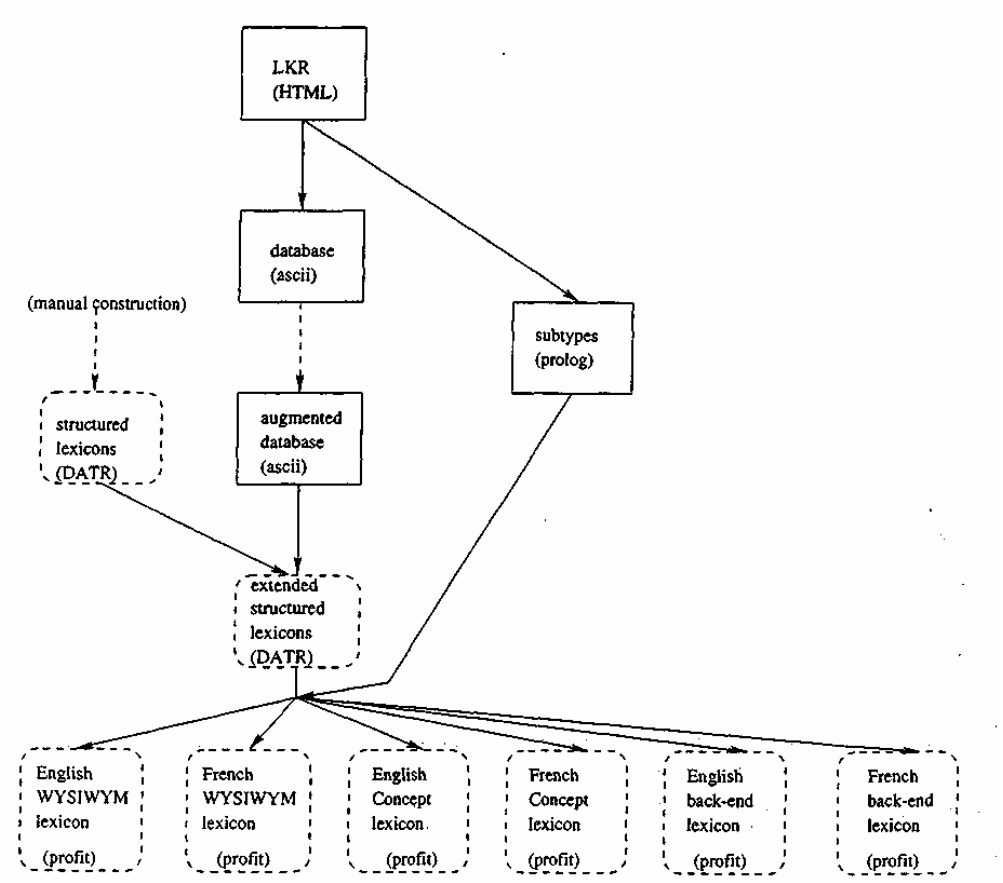

Figure 1: The automatic creation of the CLIME lexicons

The automatic extension assumed that the words could be added to existing morphological classes, so the structured morphological information at the top of the hierarchy had to be in place. The information contained in this hierarchy also had to be employed, albeit in this case in a different form, in the extension algorithm itself, since information from CELEX about the different word forms of the different lemmas was use to deduce the morphological class. Thus, for instance, in German nouns, the nominative singular, nominative plural, accusative singular, genitive singular and dative plural were all examined to infer the inflectional class of the noun. Any words which did not fit one of the classes was defined as a member of the default (regular) class and also placed in a list of entries to be checked manually.

As well as this type of monolingual deduction, the automatic extension algorithm decomposed the root forms into their syllable constituents and extracted cross linguistic commonalities across these constituents.

The semi-automatic extension of the PolyLex lexicons resulted in a fairly substantial set of lexicons for the three languages addressed. It demonstrated the use of largely unstructured databases to induce the leaves of manually constructed highly structured lexicons. However, it also has its limitations, especially form the point of view of applied NLP.

In the first place, it could be claimed that the lexicons were actually constructed from other lexicons, as the CELEX databases, although not highly structured, are nevertheless a non trivial collection of specialised linguistic data. Indeed, the availability of such sources for other languages is variable, to say the least. Secondly, the resulting lexicons themselves are probably not suitable for use in any NLP applications in their present form, due to their rather abstract nature. This suggests that we might want to consider a model of lexical construction that does not have 'input sources' and 'output lexicons' but rather a multi-layered model that my have a variety of different sources being 'refined' and combined into a variety of ultimate output lexicons.

In such a view, the PolyLex lexicons are somewhere in the middle of the layering, being a refinement of a set of already quite sophisticated lexical databases, but needing further 'refinement' to make them useable for a NLP application.

The POETIC lexicons

The POETIC project (Evans et al., 1995) was a follow-on project, further developing the TIC message understanding system to be more extendable and portable. The system takes police reports of incidents that are logged by operators and uses Information Extraction techniques to build a picture of any incidents that may affect traffic, broadcasting automatically to motorists about any relevant incidents.

The lexicon in the original system was a simple lookup table, giving syntax and semantics for each domain specific or very common English word. However, in contrast to the requirements of the CLIME NLG system described above, there was a need to have potentially several different forms for each meaning, so that all the forms that might arise in the input texts could be recognised. The revised lexicon structure, designed to simplify porting the system to a new police force sub-language, had to ultimately produce the same output as the original system. It was decided that, for these reasons, together with reasons of efficiency, the lexicons would be defined as highly structured inheritance based lexicons whose content was 'dumped' out into simple lookup tables as were in the original system. This meant that we could adapt various aspects of the lexicons, including adding quite substantial sets of new entries, relatively simply, by adding leaves to the inheritance trees. This enabled the new entries to inherit all of the more general information from higher points in the hierarchy, while the complete lexical entries required by the system were automatically generated on the basis of this hierarchically organised information.

Conclusion

We have presented a lexical architecture for NLP systems that involves potentially numerous layers of information, of possibly different granularity as well as different form. We have also presented examples of how these layers may be automatically or semi-automatically constructed. Thus, in the example of the CLIME system described above, the derivation of the monolingual database from the ontology is fully automatic. The extension of this database to be bilingual is entirely manual. The construction of the DATR lexicons from this database is fully automatic, but the next stage, to produce the ProFit lexicons, is only semi-automatic, relying on a hand-coded lexical hierarchy to be in place for the automatically derived leaves to attach to.

Essentially the same methodology was employed in the POETIC NLU system to produce lexicons for the different sub-languages used by different police forces. We believe that this

kind of approach to lexical development is the way forward, allowing the use of many and varied sources at different levels to (semi-)automatically construct, or at least extend, lexicons for genuine multilingual applications.

Note

This work was supported by the CLIME project (Computerised Legal Information Management and Explanation), EU project number EP25414. See http://www.bmtech.co.uk/clime for more details of the project. The contents of this paper were presented in a seminar at the ITRI. I am grateful to the audience for their comments on that occasion.

References

Baayen, Harald, Richard Piepenbrock & H. van Rijn (1995) *The CELEX Lexical Database* (CD-ROM), Release 2, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Briscoe, Ted, Valeria de Paiva and Ann Copestake (eds) (1993) *Inheritance, Defaults, and the Lexicon*, Cambridge: Cambridge University Press

Boguraev, Branimir and Ted Briscoe (eds) (1989) *Computational Lexicography for Natural Language Processing*, London: Longman

Cahill, Lynne (1998a) 'Automatic extension of a hierarchical multilingual lexicon', *2nd Workshop on Multilinguality in the Lexicon*, *ECAI-98*, 16-23

Cahill, Lynne (1998b) 'Lexicalisation in applied NLG systems', ITRI technical report, ITRI-99-04, obtainable via http:/lwww.itri.brighton.ac.uk/projects/rags/

Cahill, Lynne and Roger Evans (1990) 'An application of DATR: the TIC lexicon', in *Proceedings of ECAI-90*, Sweden, August 1990,120-125

Cahill, Lynne and Gerald Gazdar (1999) 'The PolyLex architecture: multilingual lexicons for related languages', in *Traitement automatique des langues*, 40:2, pp. 5-23

Copestake, Ann, Ted Briscoe, Piek Vossen, Alicia Ageno, Irene Castellon, Francesc Ribas, GerTnan

Rigau, Horacio Rodriguez and Anna Samitou (1995) 'Acquisition of lexical translation relations from MRDS', in *Journal of Machine Translation*, 9:3,1-35

Daelemans, Walter and Gerald Gazdar (eds) (1992) *Computational Linguistics* 18.2 and 18.3, special issues on inheritance

EDR (1990) Bilingual dictionary. Technical Report TR-029, Tokyo: Japanese Electronic Dictionary Research Institute Ltd

Evans, Roger and Gerald Gazdar (1996) 'DATR: A language for lexical knowledge representation', in *Computational Linguistics*, 22.2, 167-216

Evans, Roger, Robert Gaizauskas, Lynne Cahill, John Walker, Julian Richardson and Anthony Dixon (1995) 'POETIC: A system for gathering and disseminating traffic information', in *Natural Language Engineering*, 1:4, pp. 363-387

Garside, Roger, Geoffrey Leech and Anthony MeEnery (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Longman

Genelex Consortium, Report sur le multilinguisme, Version 2.0, December 1994

Hajicŏvá, Eva and Zdeněk Kirschner (1987) 'Fail-soft ('emergency') measures in a production oriented MT system', in *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, 104-108

Kameyaina, Megumi (1988) 'Atomization in grammar sharing', in *26th Annual Meeting of the Association for Computational Linguistics*, 194-203

MULTILEX (1993) Linguistic description of the MULTILEX standard, Boulogne-Billancourt: CAP GEMINI INNOVATION

Power, Richard, Donia Scott and Roger Evans (1998) 'What You See Is What You Meant: direct knowledge editing with natural language feedback', in *Proceedings of ECAI-98*, UK, August 1998, 677-68

Poznafiski, Victor, John L. Beaven and Pete Whitelock (1995) 'An efficient generation algorithm for lexicalist MT', in *33rd Annual Meeting of the Association for Computational Linguistics*, 261-267

Seamless Translation™ - Delivering for the User

by

Phil Scanlan

CEO WorldLingo

The next two years will be electric

The Internet places resources at the fingertips of users, empowering them to access information when they like, where they like and in the manner they like. This phenomenon has unleashed a demand for a paradigm shift in translation services. Enterprises that take advantage of this new paradigm will:

- Generate increased revenues
- Cut translation costs
- Achieve broader and deeper market penetration
- Provide reliable, proven services to their customer base as they enter new markets
- Be leaders in delivering a new business model of relationship services that most international customers only dream of today

The focus of this paradigm shift is not so much about the information we translate today, but the mass of information that is not translated for both economic and logistical reasons. This paradigm shift will produce winners and losers. It will change the competitive landscape forever. Winners will see an explosion in their current market share, unlock new markets, and seize opportunities. Their global competitors may never recover from the head start conceded.

Below is a clear, concise outline of the three developments key to achieving this paradigm shift. Included is also a profile of the best-suited organizations to implement, and subsequently benefit from this paradigm shift. If your organization fits the profile, a roadmap follows explaining how to leverage your translation assets in the market place to ensuring success. If your organization does not fit the profile, then you have a checklist to position your organization as a winner.

It will all be over in two years

This paradigm will not penetrate 100% of commercial enterprises and government organizations within two years. However, the early adopters of this technology will have the opportunity to discover latent needs, develop strategies to ignite these untapped resources, then gain and hold a solid advantage. Two years is all it will take to determine the winners and the losers.

How did we get here?

AltaVista's Babelfish exposed the enormous depth of demand from users for information in their own language. It has been just as effective in exposing the shortfalls in machine translation technology. Just about everybody has their AltaVista story where they translated a document into a foreign language and translated the result back again to produce a hilarious outcome. Machine translation has hit a seemingly insurmountable wall. 65% to 70% accuracy

seems to be the best that can be hoped for from machine translation as we know it today. It has remained at that level for the past 20 years. Some Machine Translation vendors claim a 90% accuracy rate, but in the eyes of a user, 90% accurate should be almost right. Users do not consider one word in every ten wrong to be almost right.

Since the translation industry could not get computers to translate accurately, they focused on helping people perform translations faster and better. On building productivity tools for people who perform translations. Translation memory tools like Trados, Star, Déjà vu, and SDLX; terminology management tools; workflow technologies; quality control systems and more. The lessons learnt using these productivity tools, the assets assembled through their use, and, the newfound power of information users, will drive this paradigm shift to Seamless Translation.

The users are in charge

The position of power possessed by users on the Internet and the competitive advantage gained by organizations that meet the demands of users will drive this change. The decision to translate no longer belongs to the organization – it is at the fingertips of users. More and more software is granting the user the power to request a translation, usually one translated by computer. Microsoft, for example, has built WorldLingo's translation service into every copy of Office XP shipped and every copy of Word 2002. If a user wants a translation – it is just a click away. Every day thousands of users download and add the WorldLingo browser translator to their web browser – so the web translates on the fly as they surf (see *http://www.worldlingo.com/products_services/browser_tool.html* as an example).

This is just the beginning. Notice when you install a new product from vendors like Microsoft and Adobe, the language you speak is one of the first items you specify. Before long, the Internet will be serving the user in their native language without any interaction required from the user. The Internet will just 'know' in what language to provide the information. It will be a seamless experience for the user. Add to this the users of Babelfish type services and very soon, you will have tens of millions of users requesting translations and computers will perform those translations.

But here is the rub: accuracy. Computer translations today are just not accurate enough. The accuracy of these translations can be improved dramatically, but generally the users do not have access to the 'translation assets' (past translations, translation memory, terminology databases, etc.) essential to improve the accuracy. Those assets are held by the organization that owns the information and are inaccessible. Those translation assets must be accessible to the user requesting a translation to improve the accuracy of computerized translations.

Who are these users?

They are almost 400 million in number. Less than 40% speak English. By 2007, Accenture predicts that the number one web language will be Chinese. Research from IDC and Forrester shows users are three to four times more likely to buy when you communicate with them in their native language. In addition, they will spend twice as long on your website. The business imperative is there to get it right. They are your clients, your prospects, your employees, your partners, your kids and communication with them does not just mean being spoken at but real two-way communication that will incorporate email, chat, and real time communication systems. The only cost effective way to provide these users with translations today is by computer. However, the translations performed by computer today are not accurate enough.

The number of worldwide e-mail mailboxes is expected to increase from 505 million in 2000 to 1.2 billion in 2005, according to International Data Corp. The volume of information on the Internet today and the amount of new information produced every day by organizations, means that there are not enough people qualified as translators on the planet to keep up with demand. In addition, most users expect to gain access to everything on the Internet instantly and for it to be free. When you throw these time and cost variables into the equation, the economics and practicalities of translating such information do not stack up today.

What users are demanding?

Users want translation to be like electricity. When you want light, you just flick a switch. Most people do not think if it was oil, coal, wind, water, or sunshine that fueled the power station to produce the electricity to switch their light on. Nor do they really ponder if it is AC, DC, or three-phase power or if a candescent or fluorescent bulb that produces the light: they just want light. Translation is the same; it should be seamless. A user just wants information they can comprehend. Most do not care if man or machine produces it, if translation memory and other tools are used, if the translators are in country or not. When they request information, they expect it in their language. Why would they want it in a language they do not speak? In short, translation should be seamless.

The next two years will see the "winners" deliver translation like electricity. It shall be seamless to the user. Here are the three key areas yielding the advances to make this possible:

1. Exposing translation assets to users of the organization's information
2. The delivery of quality translations
3. Better processing of the documents for translation.

Exposing an organization's translation assets to users

Over the past decade as the use of productivity tools for human translators has grown, most organizations (or the translation houses they have hired) have had the opportunity to amass some very valuable translation assets like:

- Existing Translations
- Translation memories,
- Terminology databases,
- Localized graphics, and
- Style guidelines.

These translation assets are very specific to their company, division, or product line. Many organizations now realize the value of these translation assets and ensure translators assign ownership of these assets as part of their service agreement. However, what is the use of owning these assets if they are never used or only used for very limited purposes? Today an organization may have legal ownership, but many could not actually lay their hands on their translation assets quickly or easily. A user would have no chance of leveraging these assets to obtain a more accurate translation. The translation assets could be stored on individual PCs in the organization, in a bunch of different departments, many are probably still in the hands of the translators. The management of these and other translation assets will become the translation issue of the decade for global organizations. Management does not mean ownership. It certainly does not mean locking them up in a secure vault where they can never be used, or only used with extreme difficulty. The key goal of managing these translation

assets is to ensure they can be used cost effectively and seamlessly by users of the organization's information to obtain the best translation available.

These translation assets are a key foundation to the provision of a seamless translation system and a high priority should be placed on their management. The development of open interfaces like TMX and TBX have made it possible to gather these disparate assets into an appropriate management system. WorldLingo's TAMS™ (Translation Asset Management System™) is nearing release and is currently undergoing beta testing. It has the following objectives:

- To ensure ownership of the translation asset
- To ensure possession of the translation asset
- To provide high availability of the translation asset to users
- To handle very large volume translation assets
- To be able to drill down to serve specialist subsets of the translation asset, and
- To allow users access to the translation asset for the purpose of having the document translated as requested, but not allowing a valuable corporate asset to be pilfered.

Since this is such a key requirement for an effective seamless translation system, if your organization would like to participate in the beta program to gain first hand experience in the effective management of translation assets, then send an email to: *whitepaper@worldlingo.com* and request a beta slot for your organization.

The delivery of quality translations

The quality of a translation is much broader than simply the correct word here or there. Today 'documents' are much more than words. Documents include graphics, sound bites, video, smart tags and much more. In the world of ecommerce, documents include amounts of money specified in various currencies. Maybe these should belong in the realm of localization rather than translation. What, however, is the difference between Translation and Localization? The simple answer is users do not care. That is more detail than they want or need to know. When they seek a document, they expect it in their language. Not just the words, but all the elements.

A picture tells a thousand words

This universal truism has been totally ignored by the machine translation industry. Pictures, diagrams, and graphics that contain text are simply not translated. Yet if a picture tells a thousand words, then wouldn't a translated diagram make an enormous difference to the comprehension of machine translated documents? An untranslated diagram is like waving a red flag at a bull. It jumps off the page and says this document has not been properly translated. A translated picture may not involve the translation of any words, but rather the substitution of a more culturally appropriate version. A seamless translation system can deliver translated pictures, diagrams, graphics, animations, videos, presentations, sound bites, etc seamlessly to the user.

Money talks

Of course it does. So why would a translated document include amounts of money in foreign currency? If a user lives in Germany, do amounts in Chinese Yuan really mean anything to the user? Who knows the exact amount they will be charged once exchange rate variations are taken into account? Yet a seamless translation system on the Internet can easily access current exchange rates for calculating the conversion. The organization could define rules to ensure price points are met. A multi-currency payment facility means the user is charged the exact amount they are quoted. This is one area where Internet based translation systems can deliver a higher quality result than traditional translations. Quality that will add dollars to your bottom line.

If you want to do actual business over the internet you should make it as easy as possible for people to purchase from you, and people are much more likely to purchase if the prices are in their own currency. The amount is meaningful to them. Moreover, multi-currency payment systems are very easy and cheap to establish today, in as little as three days you could be accepting payments in 120 different currencies (for more information see *http://www.worldlingo.com/products_services/multi-currency_creditcard_processing.html*).

Leveraging the organization's translation assets

If an organization makes its translation assets accessible to users, then the users can obtain much higher quality translations, both in terms of accuracy and consistency. These translation assets include:

- Pre-existing translations
- Translation memory
- Terminology databases
- Localized graphics, pictures, diagrams, video, sound bites and other complex elements
- Specialized machine translation engines, and
- Style guidelines.

Invisibly to the user, a seamless translation system will use these translation assets to obtain the best translation available for the user. Key to this is the organizations Translation Asset Management System (see discussion of WorldLingo TAMS™ above).

Speed is a quality issue as well

Speed is a key component of the quality of a user's Internet experience. Slow page loads cause users a great deal of stress and make the whole process tiresome. Given the chance, they will gladly go to another site that does not have the delays. Even though Seamless Translation involves a lot more intensive processing, it needs to deliver the resulting translation faster. To achieve this sophisticated caching is required.

A different way of processing documents for translation

Internet documents are three-dimensional. You navigate through them, rather than read from top to bottom. Sophisticated marketers design websites to tunnel users through to the desired end. For a translation system to be seamless it must navigate through the website with the user, taking the cookies, JavaScript, perl, CGI scripts, multimedia elements, in its stride. It must deliver the user through to the end of the tunnel designed by the marketer, say to a shopping cart and payment page. This processing system must be open and able to

communicate through XML interfaces like TMX and TBX. It should efficiently access the translation asset management systems for the organization that owns the document and use the caching systems efficiently. This processing system is mission critical and must run 24 hours a day seven days a week without fault. It must be able to scale and process very high volumes of translations, because as more and more users successfully use seamlessly translated documents, they will want to access more and more documents.

Translating economics into good business sense

In a perfect world, every organization would have all their information translated into every language, by subject experts located in each country around the world. Naturally this translation would be edited and proofread. It would be localized perfectly for each country taking into account all the cultural considerations. Transactions would be in the local currency and the translations would be constantly updated so they are always current.

All communication associated with that information would take place in the local language. Moreover, all this would translate instantly. Unfortunately, economics intrudes on the perfect world and there are few, if any organizations for which the scenario painted above is feasible. Consciously or sub-consciously, every organization makes decisions about what they will translate into which languages and what they will not. Invariably that decision is driven by the likely return on the translation investment. Seamless Translation will change the dynamics of this cost/benefit analysis by introducing a mid-point This will allow foreign language speakers to access much more of an organization's information in a usable manner increasing both sales and customer satisfaction levels.

One man's trash is another man's treasure

The value a user places on a piece of information could be vastly different to the value placed on that information by the information owner. An organization may have reams of technical support notes or user group discussions that it could not cost justify having translated by a human translator.

However, if one of those notes or discussions contains the solution to a particular problem a foreign language speaker is having, the user may be prepared to pay for a human translator – but first the user has to find the likely support notes or discussions that could provide the answer. This is an optional extra you can add to Seamless Translation. It puts more information in the hands of the user, information upon which a user can make a decision – is this worth getting human translated to get the exact solution to my problem? Often when users are searching for solutions, they feel like they are searching for a needle in a haystack. If you were searching for such a needle, would you appreciate someone saying the needle is under one of those ten pieces of straw? Seamless Translation can do that for a user.

Use real data to drive your translation expenditure

Statistics based on the actual use of the organization's information show which documents are in most demand for each language. Through Seamless Translation, an organization can understand which information they will gain the most from by having it translated by a human. Using this data generated from a seamless translation system will help the organization derive greater benefit from each dollar they spend on human translation.

Theory versus Practice

In theory, many organizations would like to centrally manage this seamless translation process and the assets associated with it. However, that is not the way the Internet was built or the way it works. The Internet is incremental by its very nature – seeing what works and adding to it. The Internet moves too fast for any five year grand plan tightly controlled from the center. To be successful, a seamless translation system must allow an organization to put the infrastructure in place for use by users of its information, but then allow the users to utilize that infrastructure in an ad hoc manner – because they will.

Seamless translation

This is what we call "Seamless Translation™" and all the building blocks are either here today or are in advanced stages of development and will ship over the next 6 to 12 months. WorldLingo is already shipping and developing many of these components. We are careful to ensure easy integration with the other pieces of the puzzle. So now, it is up to you to decide whether your organization is going to be one of the winners from the Seamless Translation revolution. Will your organization use Seamless Translation to:

- Generate increased revenues
- Cut translation costs
- Achieve broader and deeper market penetration
- Provide reliable, proven services to their customer base as they enter new markets
- Be leaders in delivering a new business model of relationship services that most international customers only dream of today?

And the winners are…

As promised in the introduction, here is the profile of the organizations best placed to be winners from Seamless Translation. They will already be heavy users of translation services and have:

- Substantial translation assets:
- Translation Memory
- Terminology databases
- Pre-existing translations
- Style guidelines
- Localized graphics, pictures, diagrams, video, sound bites and other complex elements
- Large amounts of information that cannot be translated by humans in an economically feasible way
- Clients, employees, and stakeholders that speak different languages
- Recognized the decision as to what information to translate is no longer theirs. The user will translate it, however badly, if they want to. The challenge for the organization is to make it as easy as possible for the user to obtain a good translation that casts the organization in a favorable light, and
- Stand to benefit commercially by making that information available to foreign language speakers through:
- Increased sales (people are 4 times more likely to buy if you communicate with them in their native language)

- Increased customer loyalty through better customer service - for example the provision of FAQs and support notes in the users language may let users solve many of their own problems and even prevent them in the first place
- Greater employee involvement in international policies, discussions, and project teams

If they

1. Put a strategy in place to appropriately make their translation assets available to users of their information and start to roll it out over the next 3 to 12 months
2. Start making Seamless Translation technologies available to users of their information over the next 6 to 24 months. Not by employing a big bang approach, but a piece-by-piece implementation that allows for learning and refinement.
3. Invest in translation assets that have a high leverage ratio. This might be by having the common graphical elements on the organizations website localized into several languages. Or expanding their terminology database to include the terms their organization uses over and over again.
4. Look at the business process the information is involved in and address any weak links. So if it is an ecommerce site focused on selling goods – get a multi-currency payment facility in place. If it is a customer support site, make sure you have the infrastructure for effective two-way communication in place – multi-lingual chat and email translation.
5. Recognize the information user may place a higher value on a more accurate translation than the organization does. Make it easy for the user to request a more accurate translation and leverage the organization's translation assets to make it cost effective for the user. The side benefit for the organization is the translation funded by that user adds to and enhances the organization's translation assets.

Jean Véronis (ed.) *Parallel Text Processing: Alignment and Use of Translation Corpora*
Kluwer Academic Publishers, London, Hardback, 402pp., ISBN 0792365461, Price £99

One point that is worth making clear is that this book uses the phrase 'parallel text' in the Computational Linguistics and Natural Language Processing meaning, that is, to refer to a text accompanied by its direct translation. In the translation and terminology communities the same expression means texts in the same domain written in different languages but which are not necessarily translations of each other. The most important characteristic of a parallel text (in the sense used here) is that the same thing is said in more than one language. This book provides a good coverage of contemporary methods for exploiting this feature.

This volume consists of nineteen chapters by different contributors in addition to a preface provided by Martin Kay. The first chapter is written by the editor and provides an excellent introduction to the field of parallel text processing. It covers the main methods through which parallel texts can be exploited computationally as well as describing their applications. Veronis' introduction also proves invaluable for placing he contributions in the volume within a wider context.

The remaining chapters are organised into three sections: Alignment Methodology, Applications and Resources and Evaluation. The first of these forms the core of the volume and contains nine chapters describing specific alignment techniques. There are four different levels at which a parallel text may be aligned. The most granular is at the sentence level, in which corresponding sentences in the two texts are identified. Others have attempted to align text at the level of individual words or expressions, and this is often a by-product of the sentence alignment process. Somewhere between these is the process of aligning texts at the clause level in which linguistic units somewhere between the word, or expression, and sentence are aligned. The majority of alignment methods are extended from two basic approaches reported around 1990. (Although none of these papers are reprinted in this book.) The two basic methods are lexical anchoring and sentence length correlation. The first of these makes use of the assumption that if sentences correspond then the words they contain must also correspond. So these techniques rely on using a mapping between the lexical items in the parallel texts to provide clues about sentence correspondences. In contrast, sentence length correlation methods do not focus on particular lexical items but use the observation that there is most likely a correlation between sentence lengths in the two languages being aligned. The techniques reported in this volume make use of one, or both, of these techniques. In addition some other assumptions are often employed to reduce the search space of potential correspondences to a tractable level: the order of sentence in the two texts are very close, the texts contain few (if any) additions or omissions and, finally, the majority of sentences in one language correspond to exactly one in the other. These assumptions may not hold for real world applications in which translations can be structured differently from the original text with, for example, tables and figures being positioned differently. This leads to a further level of alignment at the document structure level and two chapters describe approaches to this problem. One uses techniques borrowed from Cross Language Information Retrieval by treating the set of sentences in one language as a corpus and the other as a set of queries. The other approach is to make the structure of the texts clearer using mark-up languages such as XML.

The second section contains five chapters on applications of alignment techniques. The most relevant for those interested in translation are likely to be Gaussier, Hull and Ait-Mokhtar's 'Term Alignment in use: Machine-Aided human translation'. This describes work carried out at Xerox's French research centre which shows that word and term alignment methods can be used to build tools to assist human translators by providing them with examples of previously translated sentences similar to the material that has to be translated. This sort of technology is most often used within translators' workbenches. The next chapter describes experiments on cross-language information retrieval using a bilingual training corpus. It was found that a simple technique for automatically extracting a bilingual dictionary from parallel text was the most effective.

Other applications covered in the second section include lexicography, in which the parallel texts are used as evidence for dictionary creation; bilingual terminology extraction, where the translation of more complex linguistic units such as collocations and expressions are extracted from parallel texts; computer assisted language learning, a survey chapter describes how aligned texts can be used to provide a valuable resource for language learning. There are other applications of parallel text, which are not represented by chapters in his volume, one example being word sense disambiguation. Work in this area at IBM and AT&T in the early 1990s made use of the translations of ambiguous words in an aligned corpus to define word senses at the level of granularity most appropriate for machine translation.

The final section consists of four chapters concerned with resources and evaluation. Two chapters describe projects in which bilingual corpora were created. The first is a Japanese-English corpus and the second an English-Panjabi corpus. The next chapter describes the TMX (Translation Memory eXchange) format. This is an XML standard developed within the Localisation Industry Standards Association (LISA). The standard was developed by a consortium of companies and users to create a standard method for communication between the proprietary formats used in MT and related systems. The volume returns full circle with a final chapter, like the introduction, written by Veronis. In this chapter he describes the ARCADE project that was designed to provide standard methods for the evaluation and comparison of sentence alignment algorithms (and later extended to include word alignment algorithms). Like many areas of natural language processing the field of parallel text processing suffers from a lack of standard benchmarking resources. The ARCADE project is an attempt to resolve this problem for alignment techniques.

To summarise, this volume represents a useful overview of contemporary work on parallel text processing covering many aspects of this area including techniques, evaluation, standards and applications. Parallel text processing is relatively new in NLP but this volume is still quite modern since it does not contain any papers describing the early work carried out at Xerox and AT&T. However, the reader will find a collection that provides a valuable introduction to contemporary work on processing parallel texts.

Mark Stevenson, Research and Standards Group, Reuters, 85 Fleet Street, London EC4P 4AJ
Email: mark.stevenson@reuters.com

Conferences and Workshops

The following is a list of recent (i.e. since the last edition of the MTR) and forthcoming conferences and workshops. E-mail addresses and websites are given where known.

10-11 January 2001
Fourth Annual CLUK Research Colloquium
University of Sheffield, UK
http://www.dcs.shef.ac.uk/~njw/CLUK4

10-12 January 2001
IWCS-4: Fourth International Workshop on Computational Semantics
Tilburg, The Netherlands
http://www.sigsem.org/iwcs4.html

15-17 March 2001
CUNY 2001: 14th Annual Meeting of the CUNY Conference on Human Sentence Processing
Philadelphia, USA
http://www.ircs.upenn.edu/cuny2001

30 March – 2 April 2001
Corpus Linguistics 2001
Lancaster, UK
E-mail: mcenery@comp.lancs.ac.uk

2-7 June 2001
Language Technologies 2001: Second Meeting of the North American Chapter of the Association for Computational Linguistics
Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
http://www.cs.cmu.edu/~ref/naacl2001.html

14-16 June 2001
BI-DIALOG 2001: Fifth Workshop on the Semantics and Pragmatics of Dialogue
Bielefeld University, Germany
http://www.uni-bielefeld.de/bidialog/

18-20 June 2001
IcoS-3: Inference in Computational Semantics
Siena, Italy
http://www.cs.cmu.edu/~kohlhase/event/icos3/

27–29 June 2001
LACL 2001: Fourth International Conference on Logical Aspects of Computational Linguistics
Le Croisic, France
http://www.irisa.fr/LACL2001

6-11 July 2001

ACL-2001
Toulouse, France
http://www.irit.fr/ACTIVITES/EQ_ILPL/aclWeb/acl2001.html

8-10 August 2001
The 2001 Asian Association for Lexicography (ASIALEX) Biennial Conference
Yonsei University, Seoul, Korea
E-mail: asialex@lex.yonsei.ac.kr

20-24 August 2001
ESSLLI Workshop on Information Structure, Discourse Structure and Discourse Semantics
Helsinki, Finland
http://www.helsinki.fi/esslli

September 2001
Eighth International Conference on Translation
Langkawi, Malaysia
http://web.uum.edu.my/ict/

5-7 September 2001
RANLP-2001: Recent Advances in Natural Language Processing
Tzigov Chark, Bulgaria
http://www.lml.bas.bg/ranlp2001

7-9 September 2001
PALC 2001: Practical Applications in Language Corpora
Department of English Language, Lodz University
Tel: ++48 42 639 02 20 20, fax: ++48 42 639 02 18 20
E-mail: corpora@krysia.uni.lodz.pl

11-14 September 2001
PACLING 2001
Kitakyushu International Conference Center, Kitakyushu, Japan

18-22 September 2001
MT Summit VIII
Santiago de Compostela, Galicia, Spain
http://www.eamt.org/summitVIII/

January 2002- July 2002
Document Understanding Conference (DUC)
http://www-nlpir.nist.gov/projects/duc

13-17 March 2002
TMI 2002: 9th International Conference on Theoretical and Methodological Issues in Machine Translation
Keihanna, Japan
http://www.kecl.ntt.co.jp/events/tmi/
24-27 March 2002
HLT 2002: Human Language Technology Conference

San Diego, California, USA
http://www.hlt2002.org.

29-30 March 2002
HKTerm 2002: The Asian-Pacific Workshop on Terminology
Hong Kong

29-31 May 2002
LREC2002: 3rd International Conference on Language Resources and Evaluation
Las Palmas, Canary Islands, Spain
E-mail: choukri@elda.fr

7-12 July 2002
ACL2002: 40th Anniversary Meeting of the Association for Computational Linguistics
Philadelphia, PA, USA
http://www.acl02.org

23-28 July 2002
ALLC/ACH: New Directions in Humanities Computing
University of Tübingen, Germany
www.uni-tuebingen.de/allcach2002/

24 August - 1 September 2002
COLING2002: 19th International Conference on Computational Linguistics
Howard International House, Taipei, Taiwan
http://www.coling2002.sinica.edu.tw/

8-12 October 2002
AMTA2002: The Association for Machine Translation in the Americas
Tiburon, California,USA
http://www.amtaweb.org

MEMBERSHIP: CHANGE OF ADDRESS

If you change your address, please advise us on this form, or a copy, and send it to the following (this form can also be used to join the Group):

Mr. J.D.Wigg
BCS-NLTSG
72 Brattle Wood
Sevenoaks, Kent TN13 1QU
U.K.

Date: ...../...../.....

Name: .....
Address: .....
.....
Postal Code: ..... Country: .....
E-mail: ..... Tel.No: .....
Fax.No: .....

Note for non-members of the BCS: your name and address will be recorded on the central computer records of the British Computer Society.

Questionnaire

We would like to know more about you and your interests and would be pleased if you would complete as much of the following questionnaire as you wish (please delete any unwanted words).

1.
a. I am mainly interested in the computing/linguistic/user/all aspects of MT.
b. What is/was your professional subject? .....
c. What is your native language? .....
d. What other languages are you interested in? .....
e. Which computer languages (if any) have you used? .....

2. What information in this Review or any previous Review have you found:
a. interesting? Date .....
.....
.....
b. useful (i.e. some action was taken on it)? Date .....
.....
.....

3. Is there anything else you would like to hear about or think we should publish in the *MT Review*?
.....
.....
.....
.....

4. Would you be interested in contributing to the Group by,

a. Reviewing MT books and/or MT/multilingual software
b. Researching/listing/reviewing public domain MT and MNLP software .....
c. Designing/writing/reviewing MT/MNLP application software .....
d. Designing/writing/reviewing general purpose (non-application specific) MNLP .....
procedures/functions for use in MT and MNLP programming .....
e. Any other suggestions? .....
.....
.....
.....

Thank you for your time and assistance.