[*Cultivate Interactive*, issue 2, 16 October 2000]

The Soldiers are in the Coffee – An Introduction to Machine Translation

By Marieke Napier - October 2000

Research into Machine Translation (MT) has already celebrated its fiftieth birthday, yet understanding of its successes and failures is still minimal. Even the increase in availability of Machine Translation software due to the globalisation of the Internet has had little impact. User's knowledge of the complexities behind translating remains limited and judgements are based on one off personal experiences. This article aims to bring forth into the arena some of the questions behind Machine Translation and the issues that surround them. Understanding these particular questions is the only way researchers can move closer to their dream of a society no longer hindered by language barriers.

- -

What is Machine Translation?

The European Association for Machine Translation gives the following definition for MT: "*Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another"* [1]. It is the automation of translation.

One of MT's main problems is that such a definition, although correct, oversimplifies the process involved. This oversimplification has been a constant thorn in MT's side causing its defeats to appear even greater when contrasted against people's high expectations.

The type of MT which this article will concentrate on is that of the 'text to text' variety. MT can be divided into two types, Unassisted MT and Assisted MT. Unassisted MT takes pieces of text and translates them into output for immediate use with no human involvement. The result is unpolished text and gives only a gist of the source, hence the term 'gisting'. The ultimate aim of this type of MT is sometimes known as Fully Automatic High Quality Translation (FAHQT), perfect translation created solely by a computer. Examples of this form of MT include IBM alphaworks native search [2], Babel Fish 2020 [3], Worldlingo [4] and Dragon systems [5].

Assisted MT uses a human translator to clean up after, and sometimes before, translation in order to get better quality results. Usually the process is improved by limiting the vocabulary through use of a dictionary and the types of sentences/grammar allowed. The use of a 'controlled language' has been fairly successful. Some systems have also been set up to learn from corrections. Assisted MT can be divided into Human Aided Machine Translation (HAMT), a machine that uses human help, and Machine Aided Human Translation (MAHT), a human that uses machine help. Computer Aided Translation (CAT) is a more recent form of MAHT. Another area of MT that is worth mentioning here is Natural Language Processing (NLP). NLP parses sentences and determines their underlying meaning in order for databases to answer SQL queries entered in the form of a question. For further information on the structure of MT systems see the recent special report on the future of translation featured in Wired magazine [6].

The structure of MT systems can vary but all use some sort of transfer component. This component is specialised so that a pair of languages can produce a target sentence. The transfer component has a correspondence lexicon, which is a comprehensive list of the source-language patterns and phrases mapped to a target language. Some MT systems use systematic transfer systems, which apply software parsers to analyse the source language sentences. This type of transfer system means that for every two languages that translation is required between a new a correspondence lexicon must be created.

An alternate to the transfer component is an Interlingua, a type of intermediate language. A translation is made from the source language into the Interlingua and then into the target language. The benefits of using an Interlingua are that only one part is required for each language and therefore further languages can be added easily. See figure 1. Unfortunately the majority of work to date relies on comparative information about the specific pair languages. Arnold et al argue that use of an Interlingua could cause loss of information. They believe "*different languages 'carve the world up' differently, so settling the choices of vocabulary for the Interlingua will involve either (i) some apparently arbitrary decisions about which language's conceptualisation to take as basic, or (ii) 'multiplying out' all the distinctions found in any language*" [7]. Despite these problem areas a number of organisations have used this transfer method in building systems including ATLAS II by Fujitsu [8], and the Distributed Language Translation (DLT) system by Buro voor Systeemontwikkeling (BSO), a Dutch company, which uses Esperanto as its Interlingua [9].

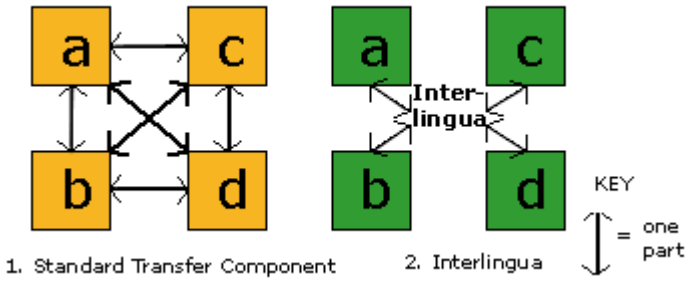

Figure 1: A comparison of a Standard Transfer Component versus an Interlingua

Why is MT Important?

You may already be wondering why MT is so important, and as this article carries on and the difficulties with MT are discussed you may argue that research into MT does not make economic sense. There are people who would agree with you. However before making such a judgement it is important to understand the underlying issues surrounding language and MT which have interested mankind for many years.

Many researchers are interested in MT because of the philosophical questions it touches upon. Each culture and religion has a different answer behind why we do not speak the same language. The Old Testament of the Bible used by the Christian religion tells how originally "*The whole earth was of one language, and of one speech*" but the people in the world were greedy and wanted everything. They even wanted to be able to reach heaven and so they built a tower up as high as they could. Naturally God caught them and was very angry. "*So the Lord scattered them abroad from thence upon the face of all the earth: and they left off to build the city. Therefore is the name of it called Babel; because the Lord did there confound the language of all the earth: and from thence did the Lord scatter them abroad upon the face of all the earth*" [10]. It is clear that most religions and cultures view Language as a strange and curious phenomenon. This has led in the past to questions such as is there a universal

language? If so is this language still around today, for example in the form of Hebrew? Is the multiplicity of language a bad thing that should be overcome? Is language integral to a people and culture? Is English the new Interlingua or cyberlanguage, after all this Web magazine, like so many others, is in English? Or is this new universal language Esperanto? These questions and some attempts to answer them are given on the Language Futures Web site [11]. MT also prompts researchers to ask whether in the future it will be possible to automate human knowledge? Will computers ever be able to think? There is no denying these are all interesting questions and to be able to answer them would not only make you rich but very famous. The philosophical implications of this type of language study also effect a whole range of sciences. MT impacts on at the very least computer science, Natural Language Processing, artificial intelligence, neural networks, and linguistics.

Pieter Brueghel the Elder (1525/30-1569 Brussels)
The Tower of Babel, Wood, H 114cm, W 155cm, Inv.no. 1026.
Image reprinted courtesy of Kunsthistorisches Museum, Vienna
URL: <http://www.khm.at>

On a more practical level there are also important political factors in the search for good quality MT. Those of us who live in places where more than one language is spoken will understand the importance of translation. There are issues, in Europe especially, over whether it is politically correct to continually use one language as the master tongue. This dominance of one language, notably English in the western world, puts other language speakers at a disadvantage. Is this dominance a bad thing? Will the enforcement of one language have propagandistic implications by not allowing people to express themselves fully? For example in George Orwell's *1984* with the use of Newspeak [12]. Does a country's loss of language in turn mean their loss of a culture? The European perspective is discussed later on in this article but it is fair to say that in Europe multilinguism is a fact of life, which makes translation necessary for communication. However translation is time consuming and continues to be expensive so MT could be a financial blessing. Reports show that soon the number of non-English web sites will out number the new sites in English. The total number of non-English-speaking online users is currently 165 million people; this will soon surpass the number of English speakers online worldwide and grow to two-thirds of the online world by 2005 [13]. There is definitely a growing financial need for some form of automated translation.

It is probably worth mentioning here that there is a fear that MT will result in millions of translators being left without jobs. In reality this is unlikely to happen. Most MT systems allow draft translation to be done then passed on to professional translators. Batch draft translation can be tedious to do anyway and translators will be given more time to spend on the more interesting intricacies of their trade.

Why can't we have MT now?

As mentioned previously, the reality is that achieving MT is difficult. We only need to look at English, a language of puns, innuendo and double entendre, to see that there are many factors that make automated translation arduous. A really comprehensive assessment of the issues behind MT and popular misconceptions is given in Arnold et al's text 'Machine Translation: An Introductory Guide' [7]. Firstly there are words with multiple or ambiguous meaning (e.g. light); then there are sentences with complex grammatical structures; and finally, to make matters worse, there are idioms. These factors all contribute to mistranslation so that we often end up with sentences as odd as the title of this article. The title is a French to English mistranslation of "*Les soldats sont dans le café*" which should actually read "*The Soldiers are in the Café*". This type of mistranslation, the French word café means both coffee and café in English, are amusing yet very easy to make. For more examples of bad translations have a look at the Fortune City Web site [14]. The article writers at Fortune City believe so strongly that accurate MT will only ever be a pipe dream that they have renamed it "*Mad Translation*" [15].

Other researchers although aware of the problems are more optimistic. Arnold et al believe that to counteract the problems MT systems will face they need to have three types of knowledge [7]. Distinctions between the 3 is not always clear but they can be defined as:

- **Linguisitic Knowledge independent of context (semantics)** - One of the ways researchers have found of dealing with the problem of semantics is by associating words with semantic features which in turn allows us to impose constraints on what other kinds of words they appear with. For example 'eat' is usually associated with edible objects. Of course there are lots of exceptions such as similes and words not used in a literal sense. Instead of using restrictions to eliminate sentences the MT system could use them to state preferences.

- **Linguisitic Knowledge that relates to context, sometimes called pragmatic knowledge (pragmatics)** - The area of pragmatics is being dealt with in many different ways. One of these is by learning the notion of focus in a sentence.

- **Common sense/real world knowledge (non-liguistic)**

The first two problems here are to do with language however the main problem MT has is not a linguistic one. Good MT is more than a system containing a bilingual dictionary and knowledge of grammar; it is more than word substitution. The main problem for MT is that computers lack real world knowledge. They do not understand the relationships things have with each other or how things fit together. For example a computer will not know that a house is bigger than a telephone, or that April can also be a female name in English. As Arnold et al explain: "*Arming a computer with knowledge about syntax, without at the same time telling it something about meaning can be a dangerous thing*" [7]. It is giving computers this knowledge that has stumped not just the MT theorists but all those working in the fields of Neural Networks and Artificial Intelligence for many years.

In his article, 'Why Can't a Computer Translate More Like a Person?' Melby calls this missing factor 'agency'. He defines agency as "t*he capacity to make real choices by exercising our will, ethical choices for which we are responsible*" [16]. Agency is closely linked to the ability to create meaning. It is almost as if MT sees as a growing child does; without insight, experience and knowledge. One way of dealing with the problem has been use of a semantic net to show relations of things to each other. Words are kept in interlinked groups that relate to other groups. How these groups relate to each other is defined by an

external source. This basic concept is similar to the idea of model groups within a mark up language like Standard General Markup Language (SGML).

This lack of understanding of the world means that MT does not work well on literary works. However, such writing is only a small percentage of the translation that needs to be done. MT can work well on text where no world knowledge is needed and there is a controlled language. Even Melby agrees that "*On some texts, particularly highly technical texts treating a very narrow topic in a rather dry and monotonous style, computers sometimes do quite well*". This controlled language can take the form of a dictionary or thesaurus or even a series of language specifications such as PACE used by the UK Company Perkins engines [17]. PACE specified that users must keep it short and simple, omit redundant words, order sentences logically, make it explicit and use the set dictionary. A number of MT systems using controlled language have been successful and remain in use today. The most cited of these is METEO, based at the Canadian Netreorological Center in Dorval, Montreal. METEO has translated around 45,000 words of weather bulletins everyday since it's main launch in 1990. The SYSTRAN system [18] has also had reasonable success and is the translation engine behind BabelFish. A list of well-known MT packages is given at About.com [19] and at Rivendel.com [20].

It seems that to work at this stage of the game MT systems need to be customised for specific purposes. At the moment good quality translation can really only be obtained by reducing what you are allowed to say.

History

The inspiration for a machine that translated from one language to another came from the code cracking of World War II. A memo drafted by Warren Weaver in 1949 contained the following lines:

"I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need do is strip off the code in order to retrieve the information contained in the text" [21].

Weaver was the vice-president of the Rockefeller Foundation and director of its Natural Sciences Division. He was an exceptional mathematician; his text co-authored with Claude E. Shannon entitled 'The Mathematical Theory of Communication' is still used in Universities today. As well as mathematics Weaver was also interested in the complexities of language and how language related to science. At a talk given in Seattle, 1966, Warren Weaver asked students not to overestimate science. *"Do not think that science is all that there is, do not concentrate so completely on science that there's nobody in this room who is going to spend the next seven days without reading some poetry…. Because if you do not open your minds and open your activities to this range of things, you are going to lead too narrow a life"* [21].

Weaver wanted a way to apply his knowledge of science to the world of rhetoric and language. In 1952 the first MT conference was organised by Yehoshua Bar-Hillel, yet by 1960 the same man published a report arguing that fully automated MT was impossible. Initially a great deal was put into MT research but as the promises failed to be fulfilled this money began to be withdrawn.

Nirenburg in his article in 'Knowledge and choices in Machine Translation' stated that "*the principal mistake of the early MT workers was that of judgement: the complexity of the conceptual problem of natural language understanding was underestimated"* [22]. This lack of judgement cost them dearly. In 1964 the National Academy of Sciences of the United

States published the report of its Automatic Language Processing Advisory Committee (ALPAC). The report recommended that most research into MT be stopped immediately due to its failure to produce useful translation. Morale in the MT field plummeted to an all time low and confidence was lost in the vision of building a fully automated MT system. Arnold et al believe that the damning content of this report can only be understood in the"'"*anglo-centric context of cold-war America where the main reason to translate was to gain intelligence about Soviet activity*". Whatever the reason for the conclusions arrived at in the report the outcome was a decrease in research in the MT field and a virtual end to all US government funding.

Over the 70s and early 80s researchers moved away from MT and concentrated on understanding language processing. A number of LP systems were set up. In the 1980s there was a revival of MT research, though much of the work was carried out in Japan and not Europe or the US. In Europe the Commission of the European Communities (CEC) did invest in the English-French version of the SYSTRAN. The moves forward that were made were seen as having "*had less to do with advances in linguistics and software technology or with the greater size and speed of computers than with a better appreciation of special situations where ingenuity might make a limited success of rudimentary MT"* [23]. Even as late as 1995 the METAL project run by Siemens was stopped after more than 30 years of development. Organisations were and still are reluctant to spend money in this area.

In 1997 Altavista adopted BabelFish Translation making it the first real time Systran translation to appear on the Web. Maybe the Internet will the catalyst that MT needs?

The European Union's Perspective

In 1958, the Council of the now the European Union declared its first regulation within which it stated that "*the official languages of the Member States should be both the official languages of the Community and the working languages of the Community institutions"* [24].

The European Union still stick by this belief and one of the main aims of *Key Action III,* of which *Digital Heritage and Cultural Content* is a key thematic areas for research, is to enable linguistic and cultural diversity [25]. It seems fairly concurrent that while the European Union is hoping for one currency it is not promoting one language. A news item about the EU's human translator service on the EU Web site states that "*Multilingualism is the cornerstone of the EU's democratic credentials. The right of every citizen to be informed and to be heard in his own language is at the very heart of democracy*" [26].

At the moment the EU translation service use the Euro-Dictionnaire Automatique (Eurodicautom), a multilingual EU-terminology database, as a translation aid. This database continues to be updated with new terms all the time. The translation service also use Systran software to a certain degree and the *TWB (Translator's Work Bench)/Euramis* System that holds previously translated pages to ensure work is not duplicated. Collette Flesch from the Translation service states in a recorded interview that while IT has opened up new horizons for multilingualism "*the Commission's needs are more at the high-end spectrum of language services, where only human translators can deliver the quality required. [However] Informatics allows the human translator to get translations done better and quicker. He can concentrate on tasks that add value, thereby leaving routine work to the machine*" [26].

Since the idea of MT arose the EU or EC as it was then has invested a fair amount of effort in MT research. It views pro language engineering as a means for communication between people when there is no common language. This positive stance for language Engineering is covered in the booklet "Harnessing the power of language".

One of the more successful projects to be funded was the European Translation Programme (EUROTRA) Project, a second Framework research and development project carried out between 1982 and 1993 to create machine translation system of advanced design [27]. The aim of the programme was to create a prototype MT system for use between the primary languages of the European Community, Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish. The projects main achievement was in the form of detailed and formalised linguistic specification. These specifications were initially going to be lost but have been maintained through the Linguistic Specifications for Future Industrial Standards Project. This project aimed to bring together information from further projects and make it available for use.

More recently during the Fourth Framework Programme (FP4) the European Commission funded a number of Language Engineering (LE) initiatives, some of which are still running. These include ELSNET [28], the European Network of Excellence in Human Language Technologies, which has been funded by the European Commission's ESPRIT and Language Engineering Programme for the last ten years. Its main role is as a co-ordinator of Human Language Technology groups. As part of the User-friendly Information Society initiative, the largest programme in the Fifth Framework Programme (FP5) LE has become Human Language Technologies (HLT) and there have been a number of significant changes. These changes are detailed on the HLT Central Web site [29]. There is more intention to fund Human Language Technologies activities that contribute to 'enhancing usability and accessibility of digital content and services while supporting linguistic diversity in Europe'. Key Action III (KA III) on Multimedia Contents and Tools specifies in its description that it aims to "*improve the functionality, usability and acceptability of future information products and services, to enable linguistic and cultural diversity. The work will address both applications-oriented research, focusing on publishing, audio-visual, culture and education and training, and generic research in language and content technologies*" [30]. A main area of this is Human Language Technologies. A full list of projects running is available from the Human Language Technologies Web site [31]. Other areas of interest include the adding of multilinguality to new applications such as subject gateways. An article appeared in *Exploit Interactive* issue 3 that considered how subject gateways can address the language needs of their audiences [32].

Another useful organisation working in the MT field from a European angle is the non-profit European Association for Machine Translation organisation (EAMT) [33]. The EAMT is a third part of a jigsaw that make up the International Association for Machine Translation (IAMT), the other two are Association for Machine Translation in the Americas (AMTA) and the Asian-Pacific Association for Machine Translation (AAMT).

The Future

The future of MT remains uncertain but with the growth of international trade and the continuing increase in use of MT technologies on the Web [34] things are looking up. More MT products are coming to market than ever before and a larger number of languages are being tackled.

Recently research has moved into the area of example-based machine translation. This method uses correct translations as a principal source of information for the creation of new ones. There is also a move towards knowledge based MT led by Carnegie Mellon University [35], and the Center for Research in Language at New Mexico State University [36]. Researchers are also continuing to investigate use and creation of an Interlingua, though the current trend is for hybridisations of a number of different techniques. However, among the

pessimistic MT professionals there remains a belief that "*research in machine translation has developed traditional patterns which will clearly have to be broken if any real progress is to be made*" [37]. Some still feel that the idea of a fully automated MT system is unachievable.

For some interesting predictions for the next fifty years in the MT world see Wired magazine's 'Machine Translation's Past and Future' [38].

Conclusion

So after fifty years of research the true art of good quality Machine Translation still remains a mystery. It seems likely though that if man can work out how to get to the moon, create a computer and split the atom then the secret of MT will someday be broken. So will the day of the Babel Fish be soon, and will it solve the answers to the philosophical, political, commercial and scientific questions it raises? When it arrives will MT be all we have hoped for? Douglas Robinson puts it well. He asks whether a machine translation system that can equal the work of a human might not be what we really want. Maybe such a MT will "wake up some morning feeling more like watching a Charlie Chaplin movie than translating a weather report or a business letter" [39].

If you are interested in using MT to translate Web pages then see Brian Kelly's article on how to extend your browser with an automated page translation feature [40].

References

1. *What is Machine Translation?,* The European Association for Machine Translation (EAMT) URL: <http://www.lim.nl/eamt/mt.html>

2. *Alphaworks.* URL: <http://www.alphaworks.ibm.com/>

3. *Babelfish 2020*, a new version with Russian-to-English, German-to-French and French-to-German translations, updated user interface and a "virtual international keyboard" URL: <http://babel.altavista.com/>

4. *Worldlingo.* URL: <http://www.worldlingo.com/>

5. *Dragon systems.* URL: <http://www.dragonsys.com/>

6. *Hello World – A Wired Special report on the future of Translation*, Steve Silberman, Wired, May 2000. URL: <http://www.wired.com/wired/archive/8.05/tpintro.html>

7. Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L. Sadler, L. (1995)Machine Translation: An Introductory Guide, Essex. Also available on-line *Machine Translation: An Introductory Guide*, Arnold and others, URL: <http://clwww.essex.ac.uk/~doug/book/book.html>

8. Fujitsu's ATLAS Machine Translation Service, Fujitsu. URL: <http://www.fujitsu.co.jp/hypertext/news/1996/Apr/23-e.html>

9. *Multilingual Machine Translation*, Dan Maxwell, Esperantic Studies, Number 3 Summer 1992. URL: <http://infoweb.magi.com/~mfettes/es3.html>

10. The Bible, Genesis 11:1-9

11. *Language futures Europe Na cànain Eòrpach san àm ri teachd.* URL: <http://web.inter.nl.net/users/Paul.Treanor/eulang.html>

12. Orwell, G. (1949) *1984*, Penguin.

13. Develop International Sales through the Internet, Global reach. URL: <http://www.glreach.com/globstats/>

14. *Examples of nice translations made by automatic translators available on the market,* Fortune City Web site: URL: <http://www.fortunecity.com/business/reception/19/mtex.htm>

15. About Machine Translation (MT), Fortune City Web site URL: <http://www.fortunecity.com/business/reception/19/index.html>

16. *Why Can't a Computer Translate More Like a Person?*, Alan K. Melby, Fortune City URL: <http://www.fortunecity.com/business/reception/19/akmelby.htm>

17. Sublanguage in the sky, *the Language Technology Group (LTG).* URL: <http://www.ltg.ed.ac.uk/papers/control.html>

18. *Systran.* URL: <http://www.systransoft.com/>

19. Machine Translation. URL: <http://ai.about.com/compute/ai/library/weekly/aa031300a.htm>

20. Machine Translation Software. URL: <http://rivendel.com/~ric/resources/mtad.html>

21. Weaver, Warren (1949) Translation in William N. Locke & A. Donald Booth (eds) (1955) *Machine Translation of Languages: Fourteen Essays*, The Technology Press of the Massachussetts Institute of Technology/John Wiley (New York)/Clapham & Hall (London), 1955. pages 15-23.

22. Nirenburg, S. (ed) (1987) *Knowledge and choices in machine Translation*, Machine Translation: Theoretical and methodological issues, Cambridge University Press.

23. *Machine Translation*, Martin Kay, Xerox-PARC. URL: <http://www.lsadc.org/Kay.html>

24. *Multilingualism, linchpin of the European Union.* URL: <http://europa.eu.int/comm/translation/en/enintro.html>

25. *The Digital Heritage and Cultural Content Web Site*, Digital Heritage and Cultural Content Web Site Development Team, Exploit Interactive, issue 2, 20 July 1999 URL: <http://www.exploit-lib.org/issue2/digicult/>

26. *Europe's tower of Babel*, Colette Flesch talks, Eur-op news - January 1998. URL: <http://eur-op.eu.int/opnews/198/en/r385.htm>

27. *EUROTRA, Centre for Computational Linguistics, Katholieke Universiteit Leuven.* URL: <http://www.ccl.kuleuven.ac.be/about/EUROTRA.html>

28. *Elsnet.* URL: <http://www.elsnet.org/>

29. *HLT Central.* URL: <http://www.hltcentral.org/>

30. *IST KA3 Home page.* URL: <http://www.cordis.lu/ist/ka3/>

31. *Human Language Technologies Web site*. URL: <http://www2.hltcentral.org/hlt/projects/index.asp>

32. *Multilingual Provision by Subject Gateways*, Marianne Peereboom, Exploit Interactive, issue 3, October 1999. URL: <http://www.exploit-lib.org/issue3/multilingual-gateways/>

33. *EAMT Home page.* URL: <http://www.lim.nl/eamt/>

34. *Multilingualism On The Web*, Marie Lebert. URL: <http://www.ceveil.qc.ca/multieng4.htm>

35. *Center for Machine Translation*, Carnegie Mellon University. URL: <http://www.lti.cs.cmu.edu/Research/CMT-home.html>

36. *New Mexico State University.* URL: <http://www.nmsu.edu/>

37. *Machine Translation: The Disappointing Past and Present*, Martin Kay, Xerox Palo Alto Research Center, Palo Alto, California, USA. URL: <http://www.fortunecity.com/business/reception/19/xparc.htm>

38. Machine Translation's Past and Future, Wired. URL: <http://www.wired.com/wired/archive/8.05/timeline_pr.html>

39. Robinson, D. (1992) Neural networks, AI, and MT, *The ATA Chronicle* volume XXI, Number 9, October 1992, The American Translators Association.

40. *Extending Your Browser With An Automated Page Translation Feature*, Brian kelly, Exploit Interactive, issue 3, October 1999. URL: <http://www.exploit-lib.org/issue3/translation/>

Author Details

Marieke Napier
Information Officer
UKOLN
University of Bath
Bath
England
BA2 7AY

Marieke Napier is editor of *Exploit Interactive* and *Cultivate Interactive* Web magazines.