Invited talk

# TWO YEARS ONLINE: EXPERIENCES, CHALLENGES AND TRENDS

Mary Flanagan

CompuServe Applied Research
5000 Arlington Centre Blvd
Columbus, OH 43220
tel: 614-457-8600
email: mflanagan@csi.compuserve.com

## Abstract

In the two years since AMTA 94, MT has established a new role in the online environment. A number of Internet based translation services have emerged and CompuServe's online translation experiments have become successful production services. The explosion of public interest in online communication offers an unprecedented opportunity for MT to establish itself as a multilingual communication tool. But along with opportunity, there are significant risks. Online MT services must meet extraordinary demands for speed, robustness and coverage to meet the needs of translation consumers. In addition, MT's capabilities and liabilities must be communicated effectively to users or the technology risks suffering a very public failure.

## 1. Experiences

### 1.1 Machine Translation at CompuServe

The MacCIM Support and World Community forums are bulletin boards where users can post and read messages. CompuServe's near real-time translation process automatically collects messages from the forums, translates them between English and French, Spanish and German and posts the unedited MT output on corresponding target forums. Translation takes place on a continuous basis and most translated messages appear on the target forum within 2 minutes. The forums have been in operation since August 1994 and February 1995 respectively and approximately 30,000 words per day are translated on each.

CompuServe Document Translation Service (CDTS) is an online upload area where members can send files for rapid, low cost machine translation. Postediting is an optional service provided at additional cost, and is outsourced to Linguistic Systems, Inc.. All translation jobs are mailed to members' CompuServe mailboxes. Unedited MT is billed at one cent per word; postedited MT costs ten cents per word. Turnaround times vary with the size of job, but rarely exceed twenty four hours. The annual volume is approximately 10 million words per year.

In progress are projects to develop translated e-mail and chat. Translated e-mail will be a standard option mail service. Because of the potentially large volumes, mail will be machine translated only; no postediting service will be offered. Translated chat will allow CompuServe members who speak different languages to communicate live and in real time online. Planned projects include web page translation, a web-based language learning area and research and development of agent technology.

## 1.2 User Reactions

Online MT produces strong reactions from users and observers, both within and outside of the translation community. Announcements of new integrations of MT online have been met with excitement or rage, but never with indifference.

End users unfamiliar with MT technology often pass through a series of phases in learning to use online MT. The first phase is amazement; the end user discovers that a translation service is available delivering high speed translations at low cost. The novice end user rarely knows to expect less than perfect results. Once the first translations are returned, amazement often turns to dismay at the errors of grammar, style and usage typical of unedited machine translation. Most end users will then proceed to the next phase, reconsideration, in which the value of raw machine translation as an assimilation tools becomes clear. Finally a pragmatism emerges when the user has learned to effectively use the capabilities of MT despite its limitations. Based on user communications, approximately 25% of new online MT users will drop out after receiving their first translations, never fully learning the value of raw MT as an assimilation tool.

Professional translators as well are rarely neutral about online MT. After the release of CompuServe's three online MT services we received hundreds of angry e-mail messages, as well as hundreds of resumes from translators:

Example flame

> Aren't you ashamed to offer such services between 3 and 10 cents per word, when actually these translations are absolutely incomprehensible for the readers concerned? Actually, you should be paying people for letting you do such a thing...You are taking advantage of the people's ignorance: By experience, I can tell you that even the best translation software cannot translate a text (Except the meteorology software used by Environment Canada, for the meteorology field has a very restricted terminology). Only (english) speakers ignorant of foreign languages can appreciate such translations, for what is translated isn't text but words.

Example resume

> ARE YOU INTERESTED IN USING MY SERVICES: I AM A FREELANCE TRANSLATOR INTO ENGLISH. I HAVE A MASTER'S DEGREE IN TRANSLATION AND AM ALSO A FULLY QUALIFIED ENGLISH LAWYER. I HAVE EXPERIENCE OF TRANSLATING FROM FRENCH AND GERMAN FOR A SWISS MULTINATIONAL AS WELL AS FREELANCE. I DO GENERAL COMMERCIAL, CORPORATE, MEDIA, FINANCIAL, BANKING AS WELL AS LEGAL TEXTS. I CAN BE CONTACTED AT 100xxx,xxx. THANKS A LOT.

## 1.3 Online Text

Online text is characterized by great variability. It ranges from informal and highly stylized forum messages to business letters and technical texts. Most users of MT in CompuServe's forums are casual consumers of information. They are using the service during leisure time to meet and chat with people from other countries. Forum messages are often hurriedly written, or written to evoke the personality of the writer, and can contain numerous spelling, punctuation and grammar errors. Forum discussions undergo continuously changing topics, called "thread drift" so the vocabulary is generally not stable.

On the other hand, MT services which permit file upload such as CompuServe's Document Translation Service typically attract business and technical users. Manuals, business letters and product descriptions are often submitted. We recently conducted a market study that revealed that the service was used mostly for business purposes where assimilation quality translation was sufficient. The users were overwhelmingly satisfied with the quality of the translations for the cost and several large users routinely submit jobs totaling more than 10,000 words per week.

Some frequency counts derived from recently collected text data on the translated forums and the Document Translation service can help characterize the differences between the source texts in each area. The 356,000 word samples show a number of differences. Forum texts have a far greater number of first person pronouns, contractions and negatives (Figure 1), reflecting the fact that forum messages are usually conversations. The higher number of occurrences of 'do' as an auxiliary in forums is likely a result of the prevalence of questions posed in forum messages. Vocabulary variability in forums is slightly less than in CDTS as well (Figure 2) with the top 50 words constituting 41% of the total vocabulary of the forums sample, as opposed to 37% for the CDTS sample. CDTS submissions are often technical documents from a wide variety of subject areas and therefore more vocabulary variability would be expected.

|  | Forums | CDTS |
|---|---|---|
| 1st person pronouns | 10,240 | 2723 |
| am/'m | 1217 | 304 |
| definite article | 14,832 | 22,226 |
| contractions (not negs) | 687 | 219 |
| do as auxiliary | 3133 | 1366 |
| not/'nt | 3779 | 2053 |

Figure 1. Forums and CDTS Frequency Counts

| Forums |  |  | CDTS |  |  |
|---|---|---|---|---|---|
| Words in top N: |  |  | Words in top N: |  |  |
| 50 | 147,409 | 41.4070% | 50 | 132,005 | 37.0801% |
| 100 | 180,294 | 50.6444% | 100 | 155,321 | 43,6295% |
| 150 | 198,999 | 55.8986% | 150 | 169,973 | 47,7452% |
| 200 | 211,876 | 59.5157% | 200 | 180,417 | 50,6789% |

Figure 2. Vocabulary Variability

## 2. Trends

Unedited MT output is gaining increasing acceptance in the online environment. This trend is driven by several factors. First, online information and participants are increasingly multilingual, creating a need for translation. The large amounts of information available online and the limited time one can spend connected encourage an online culture which favors rapid, shallow information assimilation. MT is a natural fit in this environment because it permits fast translations of adequate quality to permit a scan for content. As evidence, in CompuServe's Document Translation Service, 85% of jobs are submitted for unedited translation, with only 15% electing the services of a professional editor. In CompuServe's online discussion groups (forums) only unedited MT is available, and members use the service as a tool for communicating with CompuServe members who speak another language.

The demand for very high speed translation is also an important trend. Online users want information quickly, and translation volumes online are potentially enormous. An online MT system must be able to cope with large volumes of text and still deliver rapid turnaround. CompuServe, for example, has integrated MT into only 3 of its more than 3000 online services.

These 3 services translate nearly 30 million words per year. The ultimate conclusion of this trend is real-time online translation. Online chat areas can serve as a testing ground for this technology, though the challenges of translating the fragmented, colloquial language of chat are formidable.

## 3. Challenges

### 3.1 User Education

Educating potential MT users is the most important challenge faced by the MT industry today. The inevitable preconception of the novice MT user is that MT can produce perfect, or at least human quality translation. It is easy for MT insiders to overestimate the knowledge of the uninitiated user. MT is increasingly visible in software stores, the media and the online world. The opportunity to increase MT usership is greater than ever, and so are the risks. MT has long suffered from hype, and claims by overzealous developers and vendors of high translation "accuracy" which are not founded on any empirical measurements have set users' expectations too high. Such claims are at least partially responsible for the return rate of 20% for retail MT software; the highest in the software industry. The natural inclination of the novice user to accept MT as a translation panacea, coupled with vendor hype makes the disappointment even more acute once users see their first raw translations. In addition, most users are reluctant to read documentation and will usually miss disclaimers about translation quality or instructions about what texts are best suited to MT. The general level of knowledge of the translation consumer needs to be raised about what to expect from Machine Translation. It is essential for the MT industry to work together not only to promote MT, but to set users' expectations at an appropriate level.

### 3.2 The Online Environment

An online translation service places relentless demands on an MT system. High speed translation is important both to handle large volumes of translation, and because users expect information to be delivered rapidly, often in the same connect session. Much of the information online has diminishing value over time; messages posted today rapidly become old news. Users want translations right away, or not at all. Systems must be robust and production worthy to be able to translate potentially hundreds of millions of words per year.

### 3.3 Tools

The quality of online MT could be improved with fully automated pre-editing tools. Since most online text is hurriedly written, there is a particular need for spelling and grammar correction to enhance the translatability of source texts. The rule-based spelling and grammar correction programs available today are not adequate for fully automatic operation as pre-editors to MT. Spelling correctors often make numerous suggestions which the user must sort through. These programs might be aided by statistical information which prioritizes the suggested spellings according to frequency of occurrence in context. For example, my spelling corrector suggests more than a dozen options for the misspelling "teh" (the), and places the words "tea", "ted" and "tee" before "the" even though "the" is the most frequent word in most English texts. Grammar correction programs focus too much on stylistic issues to be valuable as MT pre-editors. Pre-editing grammar checkers which can identify and correct agreement errors, sentences with too many clauses, ambiguous structures, punctuation errors, jargon and idiomatic expressions could make source texts more readily translatable.

### 3.4 Dictionaries

Development of specialized dictionaries rarely pays off in general online MT. The tendency toward topic drift and the large number of writers make for more vocabulary variability than in subject or organization-specific texts. MT developers should concentrate on building good general dictionaries while taking care to include stable online jargon and software and hardware terminology.

## 4. Conclusion

The online translation environment is characterized by transience of subject matter and strenuous demands for speed, robustness and coverage. In addition to the online uses of MT now available, the technology may be productively coupled with Internet search engines, web browsers and language detection programs, and new and unimagined uses of the technology will likely emerge. MT will face great challenges in meeting the demands and expectations of users and in doing so in the fast-paced online environment. The future of online MT belongs to the nimble.