

# The Use of Machine Translation by Law Librarians—A Reply to Yates\*

Harold Somers\*\*

*A recent article in Law Library Journal described an evaluation of the Babelfish online machine translation service in translating legal information. Professor Somers suggests some respects in which the evaluation and, particularly, the conclusion of the article were flawed, proposes an alternative task-oriented evaluation method, and concludes that machine translation does have some use for law librarians.*

¶1 In a recent article in this journal,<sup>1</sup> Sarah Yates describes an evaluation of the free online machine translation (MT) service Babelfish, with particular reference to its possible use by law librarians. Her article discusses translation in general, how MT works, and special problems regarding legal translation, before presenting an evaluation of Babelfish translations of ten Spanish and ten German sentences drawn from legal texts. In this reply, I wish to comment on and correct some of Ms. Yates's observations and to suggest a more rigorous way of evaluating MT use for law librarians. Although the article abstract indicates that Yates "concludes that Babel Fish [sic] is not appropriate for most uses in law libraries,"<sup>2</sup> in the text of her article, she concludes that Babelfish "can usually convey enough of the gist of a text to indicate what type of information the document contains."<sup>3</sup> I will suggest, in what follows, that the latter comment more accurately suggests a real use for free online MT for law librarians and practicing lawyers.

¶2 Yates states that Babelfish was the first MT system freely available on the Web and that it is one of the most heavily used and arguably the most well-known system.<sup>4</sup> This is essentially true, though it should be noted that Babelfish is a translation *service* hosted by AltaVista since 1997, which makes use of MT systems developed by Systran, the oldest and probably the best MT provider around,

---

\* © Harold Somers, 2007.

\*\* Professor of Language Engineering, School of Informatics, University of Manchester, Manchester, England.

1. Sarah Yates, *Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information*, 98 LAW LIBR. J. 481, 2006 LAW LIBR. J. 27.

2. *Id.* at 481.

3. *Id.* at 500, ¶ 64.

4. *Id.* at 481–82, ¶ 4 (citing Winfield Scott Bennett, *Taking the Babble out of Babel Fish*, LANGUAGE INT'L, June 2000, at 20, 20).

and that the Babelfish arrangement was predated by a collaboration starting in 1994 between CompuServe and Systran to make MT available to the former's subscribers.<sup>5</sup>

¶3 Yates gives a succinct summary of some of the problems of translation in general,<sup>6</sup> stressing that word-for-word dictionary look-up is insufficient and focusing on the differing role played by syntax from language to language.

### Machine Translation

¶4 Attention is then turned to MT. Yates first states that there are three methods of MT, namely, "direct," "transfer," and "interlingua," Systran being an example of the first.<sup>7</sup> In fact, it would be more informative nowadays to make the distinction between two competing architectures in MT system design, namely, "rule-based," in which systems operate according to a set of programmed instructions, usually developed by linguists, and "statistics-based," where translation is performed on the basis of probabilities derived automatically from large collections of previously translated text. The direct, transfer, and interlingua methods are all variants of the rule-based approach, and the distinctions between them are somewhat blurred in practice. Systran, and indeed the majority of commercially available MT systems, is a rule-based system, and owes its success to a long history of continual development within this framework.

¶5 One aspect of MT so obvious that it hardly needs stating is that the quality of the output depends crucially on the nature of the input. Two factors affect this relationship. The first, alluded to by Yates,<sup>8</sup> is the problem of ambiguity.

¶6 Natural language is massively ambiguous, both at a *lexical* level (where the "same" word can have a variety of meanings, and usually therefore translations) and at the *syntactic* level (where a sequence of words can have different interpretations). Yates gives the example of *tear* (the by-product of crying) vs. *tear* (rip).<sup>9</sup>

¶7 Syntactic ambiguity can be illustrated by the often-quoted pair of sentences "Time flies like an arrow" and "Fruit flies like a banana," though more subtle ambiguities can occur due to *attachment ambiguity*, as exemplified by the phrase "investigation of techniques of stress analysis by thermal emission."<sup>10</sup> In this example, it is unclear whether "thermal emission" is a method of analysis or of

5. Mary Flanagan, *Two Years Online: Experiences, Challenges and Trends*, in EXPANDING MT HORIZONS: PROCEEDINGS OF THE SECOND CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS 192 (1996).

6. Yates, *supra* note 1, at 482–83, ¶¶ 7–12.

7. *Id.* at 484, ¶ 13.

8. *See id.* at 484–85, ¶ 17.

9. *Id.* In this example, the two words happen to be pronounced differently, but this is by no means always the case with homonyms.

10. Example from W. JOHN HUTCHINS & HAROLD L. SOMERS, AN INTRODUCTION TO MACHINE TRANSLATION 94 (1992).

investigation, and whether it is techniques for analysis or analysis of techniques (compare “techniques of stress analysis” with “loss of livelihood cases,” which does not mean that cases are lost).

¶8 Usually the correct interpretation is instantly recognised by humans, due to the requirement for text to make sense. In fact, humans usually do not even notice that alternative interpretations are possible. But computers do not have access to “common sense” and so are prone to misinterpretation, a major source of mis-translation.

¶9 The second factor is the extent to which the language in the input is “covered” by the system. At the lexical level, a word may be wholly missing from the MT system’s dictionary, or a possible translation of that word may be missing. In the case where a word has several possible translations, the system may not be able to make the correct choice of translation. This is particularly problematic when an MT system, such as Systran, which is designed for the most general usage, does not “know” about the specialised use of a word as a technical term. An obvious example from the legal domain is the word *case*, which in general language might be more likely to refer to a suitcase or a case of wine rather than a court decision.

¶10 Similarly, at the syntactic level, some sentence structures may be too complex for the system, or may have been thought sufficiently unusual for the system designers to choose not to handle them. In MT there is generally a trade-off between robustness, that is, the need to produce some sort of translation whatever the input, and quality, which depends on the input conforming to what the system designers expected.

¶11 All of these factors need to be taken into account by users of MT systems. The wrong choice of lexical item, a syntactic structure too closely modeled on the source text, or both can indeed lead to output “so strange as to make the sentences nearly incomprehensible.”<sup>11</sup> The question is whether, despite these shortcomings, MT can be of any use to law librarians.

### Special Problems in Legal Translation

¶12 So far so good. Yates now turns to the particular case of legal translation, noting that “[l]egal language has many peculiarities that make it particularly difficult to translate.”<sup>12</sup> Apart from assigning special meanings to everyday words, as noted earlier, legal language (both English and others) includes “long and complex sentences with unusual word order and other odd features” that “make legal language convoluted, cumbersome, and hard to comprehend.”<sup>13</sup>

---

11. Yates, *supra* note 1, at 484, ¶ 18.

12. *Id.* at 486, ¶ 21.

13. PETER M. TIERSMA, LEGAL LANGUAGE 69 (1999), *quoted in* Yates, *supra* note 1, at 486, ¶ 22.

¶13 In addition to these linguistic obstacles, Yates notes the poor correspondence between legal systems, concepts, and associated terms between languages,<sup>14</sup> leading her to ask whether “machine translation [can] meet this challenge?”<sup>15</sup>

¶14 A more reasonable and prior question, it seems to me, is whether *human* translation can meet this challenge? Yates cites various commentators and concludes that even the most basic requirement—to “give a complete transcript of the ideas of the original work”<sup>16</sup>—is impossible to fulfill for legal translation, even as done by humans.<sup>17</sup>

¶15 Yates therefore revises her question to ask “whether Babelfish accurately conveys the gist of law-oriented texts.”<sup>18</sup> This is a more reasonable aim, although interestingly Yates does not seek to define what she means by “gist” and, as we will see, her notion of “accuracy” is entirely subjective. In the next section, I will try to suggest how the evaluation reported fails to meet the standards usually expected of MT evaluations.

### Evaluation of Machine Translation

¶16 Yates correctly states that although there have been systematic evaluations of MT, none has focused on translation of legal texts. She then cites the 1999 study by Paul A. Watters and Malti Patel as being “perhaps the most well-known.”<sup>19</sup> This is unfortunate. There is an extensive literature on MT evaluation, but I had never heard of this study until seeing it cited by Yates. And on inspection I understood why. Their method, to translate four common proverbs from English into five other languages and then back again into English, would be rejected in an instant by any serious evaluator of MT.

¶17 Yates acknowledges that their evaluation “has some shortcomings.”<sup>20</sup> The technique of translating into a foreign language and back again, termed “back-and-forth translation” or “round-trip translation,” is unrealistic and can magnify translation problems, as Yates mentions. In fact, there is more to it than that.<sup>21</sup> A bad round trip might be due to a poor back translation of what was a perfectly rea-

14. Yates, *supra* note 1, at 486, ¶ 25.

15. *Id.* at 487, ¶ 26.

16. *Id.* at 487, ¶ 28 (quoting ALEXANDER FRASER TYTLER, *ESSAY ON THE PRINCIPLES OF TRANSLATION* 16 (Jeffrey F. Huntsman ed., John Benjamins B.V. 1978) (1791)).

17. *Id.* at 488, ¶ 28 (quoting John E. Joseph, *Indeterminacy, Translation and the Law*, in *TRANSLATION AND THE LAW* 13, 17 (Marshall Morris ed., 1995)).

18. *Id.* at 488, ¶ 29.

19. Yates, *supra* note 1, at 489, ¶ 33 (citing Paul A. Watters & Malti Patel, *Semantic Processing Performance of Internet Machine Translation Systems*, 9 *INTERNET RES.* 153 (1999)).

20. *Id.* at 489, ¶ 34.

21. See generally Harold Somers, *Round-trip Translation: What is It Good for?* in *PROCEEDINGS OF THE AUSTRALASIAN LANGUAGE TECHNOLOGY WORKSHOP 2005*, at 127 (2006), available at <http://www.altas.asn.au/events/altw2005/cdrom/pdf/ALTA200519.pdf>.

sonable outward translation, or vice versa, or both. Furthermore, a good round trip might mask a poor outward translation. For example, Babelfish translates “tit for tat” into the completely meaningless Portuguese phrase *melharuco para o tat* and back into “tit for tat,” giving the false impression that it can handle this phrase.<sup>22</sup>

¶18 In addition, as Yates points out, translating proverbs is a wholly inappropriate way to evaluate MT. Everyone involved in the MT world, including system designers, vendors, commentators, and observers, knows and acknowledges that MT works best with simple texts that mean what they say, and therefore they avoid figurative or syntactically unusual language. Proverbs like “a stitch in time saves nine”<sup>23</sup> are exactly the worst type of input to give to an MT system.

¶19 As already mentioned, there is a huge literature on MT evaluation. Yates cites two articles from one collection,<sup>24</sup> but there are many others, as well as conferences and workshops dedicated to the topic.<sup>25</sup> MT evaluation falls under the more general heading of software evaluation, a well-studied topic that includes international standards.<sup>26</sup>

¶20 The evaluation literature stresses the need to define clearly the purpose of an evaluation, and to be clear that the method devised fits that purpose. Application-oriented evaluations focus on issues of concern to end users (as opposed to internal evaluations for developers) such as usability, cost, and so on. In the following section, I will look more closely at Yates’s evaluation.

### Yates’s Evaluation

¶21 As Yates says, “[t]he purpose of the translation is an extremely important consideration,”<sup>27</sup> yet in her evaluation she never states precisely what purpose she has

22. *Id.* at 128.

23. One of the examples used by Watters & Patel, *supra* note 19.

24. Nico Weber, *Machine Translation, Evaluation, and Translation Quality Assessment*, in *MACHINE TRANSLATION: THEORY, APPLICATIONS, AND EVALUATION: AN ASSESSMENT OF THE STATE-OF-THE-ART* 47 (Nico Weber ed., 1998), *cited in* Yates, *supra* note 1, at 489 n.45; Rita Nübel, *MT Evaluation in Research and Industry: Two Case Studies*, in *MACHINE TRANSLATION: THEORY, APPLICATIONS, AND EVALUATION: AN ASSESSMENT OF THE STATE-OF-THE-ART*, *supra*, *cited in* Yates, *supra* note 1, at 490 n.48.

25. For example, the LREC (Language Resources and Evaluation Conference) series, the most recent having been held in Genoa, Italy, May 2006. Major publicly funded evaluation initiatives include the European Union’s EAGLES project (Expert Advisory Group on Language Engineering Standards); the related ISLE project (International Standards for Language Engineering), cofunded by the EU and the USA’s National Science Foundation and described at ISLE, <http://www.issco.unige.ch/projects/isle/ewg.html> (last visited Mar. 17, 2007); and the United States’ ARPA initiative, described in John S. White et al., *The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches*, in *TECHNOLOGY PARTNERSHIPS FOR CROSSING THE LANGUAGE BARRIER: PROCEEDINGS OF THE FIRST CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS 193 (1994)*, available at <http://www.mt-archive.info/AMTA-1994-White.pdf>.

26. For example, the ISO/IEC 14598 series. The International Organization for Standardization has a standing committee, Software Product Quality Requirements and Evaluation (SquaRE), that is also working on specifications.

27. Yates, *supra* note 1, at 491, ¶ 41.

in mind. She proposes to evaluate “the program’s ability to express the general meaning of the text,” that is, “the same general meaning as [that conveyed by] a professional translation,” where “same meaning” means “in an ordinary context, but not necessarily in a strict legal sense.”<sup>28</sup> This is rather vague.

¶22 The evaluation involves translation from Spanish and German into English of five sentences each from the Mexican and German civil codes and five sentences each taken from press releases from the foreign ministries of those two countries of interest to the American legal community. For all the texts chosen, professional translations into English were also available, which allowed Yates to make comparisons between the translations.

¶23 Interestingly, possibly the most widely used method for evaluating MT output nowadays involves comparison with professionally done translations. These comparisons are done fully automatically and involve a more or less complex calculation reflecting the degree of overlap of individual words and word sequences between the translation to be evaluated and a “gold standard” reference translation. The best known of these measures, which are so widely used as to have become almost mandatory in descriptions of experiments with MT systems, is the BLEU score developed at IBM.<sup>29</sup> The method is controversial, however—it rewards translations that use the same vocabulary as the reference translation and penalizes translations that use semantically equivalent words or phrases. For this reason, use of multiple reference translations is recommended. The method also fails to penalize translations that use the right words but lack grammaticality; thus, a high score can be achieved with a translation that is nevertheless quite hard to read and understand. Most significantly, however, the availability of automatic metrics has encouraged a culture in which evaluations involve hundreds or thousands of sentences, in contrast to evaluations (like the present one) involving human judgments, in which the volume of text that can be looked at is naturally restricted.

¶24 Yates compares the Babelfish translations with the human reference translations manually, cataloging errors as lexical or structural and rating them as “minor,” “moderate,” or “severe.”<sup>30</sup> This method is one that has been seen over the

---

28. *Id.*

29. See generally Kishore Papineni et al., *BLEU: A Method for Automatic Evaluation of Machine Translation*, in PROCEEDINGS OF THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 311 (2002).

Ms. Yates has kindly posted all her data on the web, at <http://www.law.umn.edu/uploads/images/2288/translations.pdf>, which allows us to make some comparisons. For what it is worth, the texts used in Yates’s evaluation get BLEU scores of between 0.1139 and 0.1406; a score of around 0.4 is usually considered to represent a fairly good translation, so the automatic evaluation confirms the poor quality of the Babelfish translations, although it should be stressed that the BLEU metric is usually applied to much larger volumes of translated material.

30. Yates, *supra* note 1, at 493–94, ¶¶ 46–50.

years,<sup>31</sup> usually performed by linguists and involving a quite sophisticated categorization of error types, with the result often fed back to the developers who would then try to correct the errors.

¶25 Error rating is, of course, a subjective process, and well-known problems associated with subjective judgments are usually addressed by doing things on a larger scale, that is, by having many judges making many judgments. Yates's evaluation is weak on this score: she has one other judge, but instead of having her do a parallel evaluation, she is asked only to review Yates's own ratings.<sup>32</sup>

¶26 The results of this evaluation are damaging for Babelfish. Fifteen of the twenty sentences (75%) are "considered failed translations" since they contain at least one "severe error."<sup>33</sup> But this result immediately reveals a flaw in the scoring method. The sentences are mostly very long,<sup>34</sup> but just one severe error in a sentence is sufficient to have it rated a failure, even though some high percentage of the rest of the translation might be quite comprehensible. What is more, a severe error counts the same whether or not it affects a more or less important part of the text.

¶27 For example, the first translation in the data set, from the Spanish Civil Code, reads as follows: "The spouses will contribute economically to the support of the home, its feeding and the one of their children, as well as to the education of these in the terms that the law establishes, without damage to distribute to the load in the form and proportion that decide for this effect, according to their possibilities." Yates classifies the mistranslation of *perjuicio* as "damage" rather than "restriction" as severe, and thus as rendering the entire translation to be a failure. Even if the translation from "without damage" onwards is, admittedly, rather garbled, it must surely be conceded that the translation up to that point, representing about 60% of the text, is perfectly acceptable.

¶28 Nowhere in the analysis of the results is the length of the input taken into account. For example, Yates finds that Babelfish handles German "markedly better" than Spanish,<sup>35</sup> frankly, a big surprise to anyone who knows Systran or, indeed, these two languages. Spanish sentences contain roughly twice as many errors—not surprising since they are on average exactly twice as long (average 43.6 words vs. 21.8).

31. See, e.g., Frank Knowles, *Error Analysis of Systran Output: A Suggested Criterion for "Internal" Evaluation of Translation Quality and a Possible Corrective for System Design*, in *TRANSLATING AND THE COMPUTER* 109 (Barbara Snell ed., 1979); Mary A. Flanagan, *Error Classification for MT Evaluation*, in *TECHNOLOGY PARTNERSHIPS FOR CROSSING THE LANGUAGE BARRIER*, *supra* note 25, at 65.

32. Yates, *supra* note 1, at 494, ¶ 52.

33. Yates, *supra* note 1, at 494, ¶ 53.

34. The average lengths (in words) for the test sentences are as follows: Spanish Civil Code, 30.8; German Civil Code, 27.6; Spanish press releases, 56.4; and German press releases, 16.0.

35. Yates, *supra* note 1, at 494, ¶ 55.

¶29 Yates's analysis of the distribution of minor, moderate, and severe errors and their relative frequency in the two languages and two text types would be of potential interest if there were more data, and if they were more clearly linked to some practical measure of the usability of the translations. This brings us to a proposal for a more focused evaluation.

### An Alternative Evaluation Scheme

¶30 A better evaluation method, appropriate for readers of *Law Library Journal*, would investigate fitness for purpose. So the first question we must ask is what kind of text is a law librarian likely to want to translate, and why?

¶31 Clearly, a text for which a translation already exists, such as the civil codes, does not meet this requirement. This is not to be seen as yet another criticism of Yates's study: she obviously needed a text for which there existed a translation because she wanted to compare the two.

¶32 Since we know that Babelfish will not produce a perfect translation, we need to find a scenario where a less-than-perfect translation would nevertheless still be welcome. Almost every introductory article or book on the subject stresses that MT does not produce a 100% reliable translation, and suggests that it can and should be used to get a rough idea of the content of a text, perhaps to see if it is worth getting translated "properly." This is particularly the case where the alternative (to a poor translation) is no translation at all.

¶33 One form of evaluation that is sometimes proposed is to get people to perform some task on the basis of translated material.<sup>36</sup> Care has to be taken that this method really evaluates the translation, and not their ability to perform the task *per se*, and it can be a delicate matter to quantify the results.

¶34 An evaluation design that might meet this requirement could consider a summary of a case that a lawyer might be interested in, but which is written in a foreign language. We could test the capability of Babelfish to produce an understandable translation of a case summary by setting a series of factual questions based on the original text, and asking the subject to answer the questions based on a Babelfish translation of the report. This is not dissimilar to the use of reading comprehension tests, a technique that has been proposed quite frequently over the years.<sup>37</sup>

¶35 Summarizing the important aspects of a reported case is a genuine task for trainee and practicing lawyers, and the "right answers" can be established beforehand, so that grading the replies is straightforward. In this way we would have

---

36. See, e.g., Kathryn Taylor & John White, *Predicting What MT is Good for: User Judgments and Task Performance*, in *MACHINE TRANSLATION AND THE INFORMATION SOUP* 364 (David Farwell et al. eds., 1998).

37. See Harold L. Somers & Nuria Prieto-Alvarez, *Multiple-Choice Reading Comprehension Tests for Comparative Evaluation of MT Systems*, in *WORKSHOP ON MACHINE TRANSLATION EVALUATION AMTA-2000*, at 107 (2000) (including discussion and extensive list of references).

an objective rather than subjective assessment of the fitness for task of Babelfish. What is more, because collecting the results is simple, and could even be automated if necessary, the only barrier to conducting a wide-ranging comparative assessment would be finding sufficient numbers of willing subjects. We plan to carry out such an evaluation in the near future.

### Conclusions

¶36 Yates's article concludes with a discussion of what, given the poor performance of Babelfish, are the alternatives when dealing with foreign-language legal information. She suggests three: try to find the information in English, get an amateur human translation, or pay for a professional translation.<sup>38</sup>

¶37 The suggestion to rely on amateur translation is one that professional translators are always wary of, their profession having a long history of being undervalued, though to her credit Yates suggests only that such a translation is likely to be better than Babelfish's and, like Babelfish, should not be seen as a real alternative to full translation. It would be interesting to test this theory by having the same twenty Spanish and German texts translated by someone "with some foreign-language skills."<sup>39</sup> How to quantify these skills is another matter, of course. Are we talking about two years of high-school Spanish, a German grandmother, or something a lot more sophisticated? And we should not forget that even a bilingual person, or indeed a professional translator who does not specialise in legal texts, is not likely to produce a perfect translation (whatever one of those is).

¶38 Yates firmly concludes that professional translation is the only way if the information is vital, but then concedes that "[i]ronically, this is one case where Babelfish might come in handy."<sup>40</sup> Babelfish can indeed produce a rough translation from which sufficient information can be got to indicate whether a full translation (and of which parts) is necessary. This seems to me to be a really useful service, and one that readers of *Law Library Journal* might make extensive use of. To conclude, as Yates does, that "Babelfish is of little use in a law library,"<sup>41</sup> a finding prominently repeated in the article's abstract, seems to me to belie the findings reported in the rest of her work.

---

38. Yates, *supra* note 1, at 499, ¶¶ 59–61.

39. *Id.* at 499, ¶ 60.

40. *Id.* at 500, ¶ 64.

41. *Id.* at 500, ¶ 67.