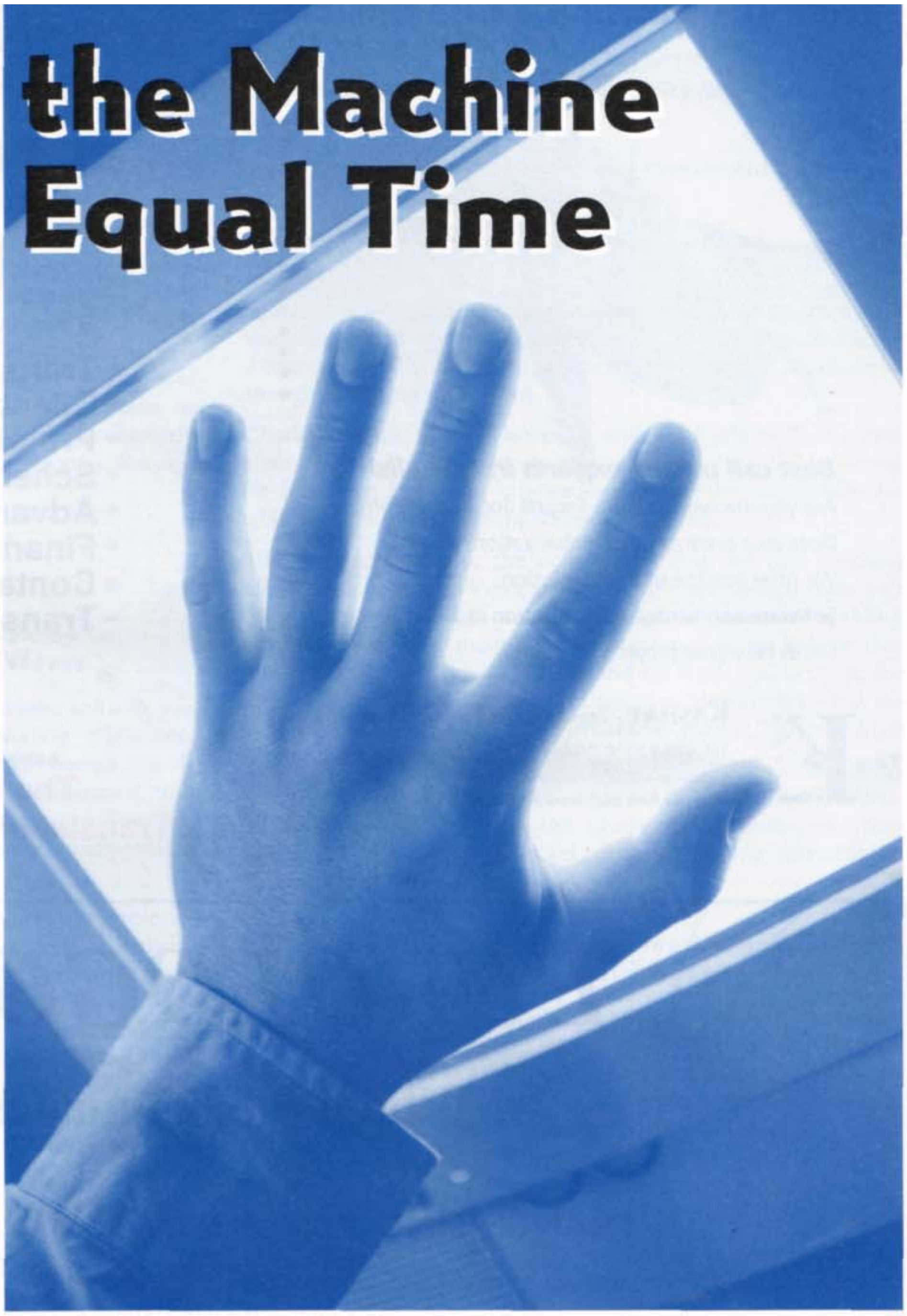
Giving the Machine Equal Time



An MT expert takes on one of machine translation's most vocal critics.

by Jeffrey Allen

In this article, I would like to comment on Douglas Hofstadter's statements that appeared in an interview article entitled "A word with Hofstadter" in *Language International* 10.1 (pp. 32–36).

I realized in reading the interview that Hofstadter had agreed to substantially comment on two fields (i.e., controlled language—CL—and machine translation—MT) that do not constitute his area of expertise; he even admits to having no experience in these fields. This is clear from his comments, including: "…I think they're not just as daunting in a sense, although maybe I would take that back if I thought about it for a while" (p. 33); "I don't think it's a solution at all" (p. 34); "that I can't compare because I haven't been in the field" (p. 34); "maybe again I haven't followed MCE [Multinational Customized English]; I haven't followed Systran, but my guess is that a lot of what comes out is probably fairly incomprehensible" (p. 34); "I don't use the Web, so I can't comment on it [automatic translation of documents via Web sites] from personal experience. It sounds ridiculous to me. It sounds absolutely preposterous" (p. 34); and "I guess that's some kind of labor saving" (p. 34).

All of the above-mentioned subjective statements appear within his interview that is composed of 15 individual replies (12 full paragraphs and three individual sentences) specifically addressing the controversial topics of CL and MT. Hundreds of articles on CL and MT have appeared in the proceedings of regular conferences of the Society for Technical Communication, the Controlled Language Applications Workshop, the MT Summit, the Association for MT in the Americas, the European Association for MT, the Association for Computational Linguistics, the European Association for Computational Linguistics, Translating and the Computer, Computational Linguistics, etc.

I will only provide a handful of references in order to demonstrate that a significantly large amount of objective data was available on CL and MT before Hofstadter's comments appeared in the interview; a few recent references are also provided. I hope that this information will help clarify some of the possible misconceptions of CL and MT that were presented in his interview.

In the first instance, I understand Hofstadter's perspective where he says that "they [Microsoft] don't bother to write them [operating systems and manuals] well in the first place, so who cares about translating them?" (p. 33). Having myself worked in and with several large corporations in the field of authoring and translation technologies, I have noticed that such organizations are now becoming more aware of the need to improve the workflow process and the quality of their multilingual technical, marketing, and internal documentation due to a higher amount of product sales in non-English-speaking countries. It is thus not surprising that much emphasis has been placed on these areas of communication over the past few years. These documentation processes are discussed in Allen (1999a, 1999b), Allen and Smart (1999), and Schubert (1999).

I also agree with Hofstadter that applying MT to professional literary translation is not worth implementing due to the highly stylistic nature of literary texts. I do not know of any academic research being conducted on the application of MT to literary texts. The only indication of something even remotely related to this is the experiment conducted by Ostler (1999) who used MT to automatically translate a philosophical text (Wittgenstein, 1921) as an interesting exercise. In addition, MT systems are not used for the translation of literary texts by freelance or translation agencies. All one needs to do is go to the online electronic archives of the Language and Translation (Lantra) List (http://segate.sunet.se/listinfo/lantral/index.html) and search on "MT" and/or "machine translation." This will provide many hits on comments made by professional literary translators with an opinion similar to that of Hofstadter. Many of my comments during April to November 1998 on that list explain the different types of MT systems, the different approaches to translating with computers, and factors to consider when deciding whether or not to use MT.

Later in the article, I noticed that Hofstadter stated: "So when I hear these claims of companies, part of me reacts with a kind of a sneer, 'Oh, cut it out, this is junk; this is nonsense!'" (p. 33) Various benchmarking and ongoing advanced tests on the application of CL and/or MT in companies and research centers (e.g., Nortel, General Motors, Caterpillar, Xerox, Center for Machine Translation at Carnegie Mellon, etc.) have demonstrated that the accuracy of MT output depends on the type of document being translated. In my own experience, helping develop technical authoring and translation systems, and then training users on them, has led to my participation in (unpublished) studies that have shown that technical procedural texts written in a CL produce better MT output than do long descriptive texts.

With regard to overgeneralized statements concerning unusable MT output, I can only respond by saying that one must face the fact that "Meteo 96's results were 16 days of near-perfect (93.2 percent) weather translations with real-time access to the target language by Info 96" (Chandioux and Grimaila, 1996, p. 5) during the 1996 Olympic Games in Atlanta. Published results on another set of CL and MT systems are also found in *Xerox: traduction automatique fidèle dans 42 mois?* (www.ledevoir.com/redaction/planete/pla_accueil.html).

I add a disclaimer that figures at over 90 percent are based on specific needs, expectations, and document types that have been well studied and analyzed for implementation in particular domains, contexts, and environments. However, these figures still show that MT output can be quite usable when applied correctly.

CL writing principles did not magically appear by asking the computer to create them. Rather, CL principles are the result of large-scale terminological and linguistic analyses based on abundant corpora and legacy texts found in archives of the organizations wishing to implement these authoring and translation solutions.

Hofstadter then says, "I really do mean clarity. And I don't think you're going to get that," and "I'm sure you can have somebody write some kind of gobbledygook in Multinational Customized English that some sort of drone can read and say, 'Yeah, this sounds okay to me!'" (p. 34) Again, I must emphasize that this statement is based on seemingly non-objective claims. Shubert et al. (1995), Mitamura and Nyberg (1995), Holmback et al. (1996), Goyvaerts (1996), Farrington (1996), Chervak et al. (1996), Van der Eijk (1997), Barthe (1998), Knops and Depoortere (1998), Mellor (1998), Allen (1999a, 1999b, 1999c), all very clearly state that improved source-language comprehensibility, as well as target-language comprehensibility, are the fundamental points for industrial and corporate investment in CL and/or MT applications.

CL writing principles did not magically appear by asking the computer to create them. Rather, CL principles are the result of large-scale terminological and linguistic analyses based on abundant corpora and legacy texts found in archives of the organizations wishing to implement these authoring and translation solutions.

Mitamura and Nyberg (1995) have described their procedures of analyzing large amounts of legacy texts. Allen and Smart (1999) have discussed how data-mining is one of the procedures for successfully implementing CL checkers into an industrial environment. It is important to note that the principles forming the core of CL system-writing rules are often enhanced principles of technical writing and technical communication optimized for automatic language processing. I even heard a presentation a few weeks ago that described an approach for pre-editing documents in a "weakly controlled language" that "does not impose any unnecessary restrictions on the writers, but rather addresses only the issues of importance to the MT process" (Bernth, 1999). Several other key papers previously presented and written on these topics can be found in the proceedings of CLAW'96, CLAW'98, AMTA'98, *Computational Linguistics in the Netherlands*, etc. For relevant lists of papers, I refer *Language International* readers to the following Web sites:

www.ccl.kuleuven.ac.be/claw/programme.html;

www.lti.cs.cmu.edu/claw98/schedule-onsite.html;

salto.let.uu.nl/www/controlled-languages/home.html/;

www.up.univ-mrs.fr/~veronis/claw2000

Although I still hear once in a while that CLs are just an academic exercise for theoretical issues in translation research, such a statement is outweighed by the steadily increasing number of industrial and corporate players seriously considering this solution for their multilingual documentation needs. I list but a few organizations involved in implementation efforts: Caterpillar, Océ, Nortel Networks, Diebold, Eastman Kodak, and Lucent Technologies. Industrial users include Rolls-Royce, Boeing, Aérospatiale, General Electric, British Airbus, GM, and Renault. Government administrations include the Dutch tax authority and the US government.

The list does not stop there. Some translation vendors and localization agencies, who certainly refrain from investing human and financial resources in research activities that do not have a clear return on investment, have also been focusing on CLs, including:

(continued on page 44)

(continued from page 25)

Trados (Brockmann, 1997), Star (Janssen et al., 1996), SimulTrans (*SimulTimes*, 4 Oct. 1999), and Lant (Knops and Depoortere, 1998). Daniel Brockmann of Trados states that, "the more controlled a source text, the more efficient these tools will be in the translation process. In the midterm, they will also be adapted for source-text authoring. This means that the writer will be able to reuse his or her own material using an *authoring memory*, thus increasing consistency even more in the source language." (Brockmann, 1997, p. 10) Lant's efforts have been fruitful as noted by their client, GM, which announced that the Controlled Automotive Service Language pilot had been "declared a total success by senior management, having met or exceeded its goals." (Controlled Languages and TCF-Gen email discussion lists, 25 Oct. 1999)

I am surprised that Hofstadter claims that Multinational Customized English (MCE) is an artificial mode of expression. He states: "in Xerox's case, we know that the stuff is fed in, in this artificial version of English..." and is "written in this very artificial English so that all ambiguities are avoided." (p. 34) First, how does Hofstadter qualify such an *artificial language* in terms of objective qualitative and quantitative data without ever having studied MCE? Second, writing rules and technical descriptions of different CLs can be found in Mitamura and Nyberg (1995), Atwater (1998), Bernth (1998), and Mitamura (1999). Both lexical (word-level) and structural (phrase- and sentence-level) ambiguity constitute a large part of the elements of these descriptions. However, the 15 writing rules of Nortel Standard English (Atwater, 1998, p. 3) are basically a mix of writing rules typically found in international technical-writing books, along with a few principles that improve MT output. All these rules clearly conform to the English language as I, a native speaker of it, know it. In another CL, the full page of examples (all in context), that are taken from a non-pre-edited text and its equivalent pre-edited text in IBM EasyEnglish (Bernth, 1999), are far from being artificial and not understandable to me.

Further, an email discussion message recently posted on the Controlled Languages and TCF-Gen lists by Linda Means of GM was written in accordance with the rules of the Global English CL which she currently teaches at GM University. Means states that Global English "is GM's corporate standard to make texts suitable for international use, in English and in translation. Global English provides ease of comprehension for non-native speakers and for human translators. GE has only 12 rules." (Means, 1999) The entire message is in very comprehensible English and is far from being artificial. Why would GM invest so many financial and human resources to teach thousands of employees to write in an artificial language for their communicative needs? I invite Hofstadter to read Beuttenmüller (1997), which contains a two-page report on the second international Controlled Language Applications Workshop. This report was then rewritten by Mark (1997) into the Didactic-Typographic Visualisation CL, and also by Farrington (1997) into AECMA Simplified English. There is no indication in either of these two rewritten versions how these CLs can be artificial languages.

I am surprised that Hofstadter claims that Multinational Customized English (MCE) is an artificial mode of expression. How does Hofstadter qualify such an artificial language in terms of objective qualitative and quantitative data without ever having studied MCE?

Rather than state that CL is artificial, it might be more appropriate to look into issues concerning mastery and proficiency in a CL that can lead to good or bad CL writing. This has already been discussed in Adolphson (1998), Mitamura (1999, section 5.2), and Allen (1999b).

Hofstadter also assumes that, "if you're in this very limited domain of discourse where the computer has been prepared with thousands of ready-made phrases so it won't stumble on those, and where you've written it in this very artificial English so that all ambiguities are avoided, then you can get output that only needs minor tweaking." (p. 34) The distinction between Controlled Language and Sublanguage can be found in well-documented existing references in the field (e.g., Grishman & Kittredge, 1986; Dachelet, 1994). What Hofstadter refers to as a "limited domain of discourse" has not necessarily resulted directly from the influence of the computer. Sublanguage is not simply the result of artificially crafted controlled input for MT, but in most cases has already been developed by human beings in a domain-specific environment prior to the introduction of technologies and computers. The advantage of subdomain work and sublanguages is that they are highly repetitive and easily translatable. It is well known that computational systems work best with existing subdomains where sublanguages have been developed. John Chandioux Experts-Conseil has shown that the subdomain of weather-report bulletins produces a sublanguage favorably predisposed to MT systems (over 93 percent MT accuracy) through the Meteo system mentioned above.

Hofstadter easily criticizes the concept of "ready-made phrases" as some type of artificially created idea. Adolphson (1998) presented a very accurate account of how technical authors have learned to write by plagiarizing, to a great extent, the texts of their co-workers. Adolphson's claims are confirmed in Allen (1999a), who shows that new memory-based technologies (e.g., authoring memory) can be adapted to this manner of learning. Plagiarizing texts, although obviously discouraged in academic circles, is highly appropriate for technical writing and is a key factor to success for implementing new authoring and translation technologies. Contrary to Hofstadter's belief, *traditional* MT systems are in fact not at all based on a *ready-made phrase* approach. He seems to be confusing the technologies of translation memory (TM) and MT. By simply mentioning *MT systems*, he is most likely referring more specifically (and technically) to the transfer-based MT approach whereby possible translation equivalents (known as "potential parses") are produced according to syntactic mapping rules. Neither this MT approach, nor the other main one used for high-quality translation (e.g., knowledge-based with an interlingua), is based on memorized stock phrases that are internally stored. Terence Lewis confirms my distinction above with his recent statement that, "It is still valid to draw a distinction between applications that attempt to analyze (parse) and translate (in some cases) a wholly unseen text and applications that are designed to compare strings in a particular language with pairs of strings in a database and retrieve complete or partial (fuzzy) matches according to various user-defined criteria." (Lewis, 1999, the former referring to MT and the latter to TM)

I add here that the only type of MT system at present that uses stock phrases is known as example-based machine translation (EBMT), which (1) is not currently used for high-quality MT needs, and (2) more closely resembles the approach taken by TM. Using such memorized segments as part of memory-based systems (i.e., translation memory, or authoring memory) has been one of the hottest growth areas in the translation and localization industry over the past five to 10 years. A full list of existing TM tools are provided in Allen (1999a). Several industrial companies are also currently adding memory-based components to their CL and MT workflow implementations (Allen, 1999b). I know many professional translators who greatly benefit from the use of TM tools, but they do not work at all with MT systems. It is therefore important to distinguish between these various translation technologies rather than lumping them all together.

In conclusion, there is overwhelming evidence contradicting the subjective and unresearched statements made by Douglas Hofstadter about CL and MT applications and their results being *artificial, unusable, incomprehensible*, etc. Hofstadter and other *Language International* readers are encouraged to consider the objective, published facts on the state-of-the-art for CL and MT. Many relevant Web sites are included in the text above and in the references below for future information.

Jeffrey Allen has been a specialist in translation, technical writing, authoring/translation-system technologies, and language teaching since 1988. He has worked as the trainer of the controlled language known as Caterpillar Technical English, a developer of machine-translation systems at the Center for Machine Translation of Carnegie Mellon University, and is currently Technical Director of text- and speech-based language-resource database projects at the European Language Resources Distribution Agency (ELDA) in Paris. He has served on several conference committees in the language technology field (CLAW2000, AMTA2000, LREC2000) and is an active member in the MT Certification and the MT Postediting special interest groups. Contact him at postediting@aol.com

References

Adolphson, Eric. 1998. "Writing Instruction and Controlled Language Applications: Panel Discussion on Standardization." In CLAW98 proceedings, p. 191.

Allen, Jeffrey. 1999a. "Adapting the Concept of Translation Memory to Authoring Memory for a Controlled Language Writing Environment." Paper presented at ASLIB-TC21.

Allen, Jeffrey. 1999b. "Implementing Controlled Language and Machine Translation in the Automotive Industry." Invited talk at the Multilingual Documentation for the Automotive Industry Toptec Symposium. Cosponsored by the Society of Automotive Engineers (SAE), the Localization Industry Standards Association (Lisa), and Alpnet. 21–22 October, Amsterdam, The Netherlands. www.praetorius1.demon.co.uk/new/toptec1.html

Allen, Jeffrey. 1999c. "Different Kinds of Controlled Languages." In *TC-Forum* magazine, volume 1–99, pp. 4–5. www.tc-forum.org/topiccl/cl15diff.htm

Allen, Jeffrey and John Smart. 1999. "English Language History, Controlled English Checkers, Writing Techniques, Examples of Controlled English." Invited talk for the annual conference of Technical Information Creation and Distribution (TICAD), 3 November 1999, Birmingham, England. www.smartny.com/download.htm

ASLIB-TC21: The 21st Conference of Translating and the Computer, sponsored by ASLIB, 10–11 November 1999, London. www.aslib.co.uk/conferences/tc21.html

Atwater, Kathleen. 1998. "Nortel Standard English as a Quality and Reliability Tool." Distributed Report. Ottawa, Canada: Public Carrier Networks Information Development, Nortel Networks, 1998.

Barthe, Kathy. 1998. "GIFAS Rationalised French: Designing One Controlled Language to Match Another." In CLAW98 proceedings, pp. 87–102.

Bernth, Arendse. 1999. "Controlling Input and Output of MT for Greater User acceptance." Paper presented at ASLIB-TC21.

Bernth, Arendse. 1998. "EasyEnglish: Addressing Structural Ambiguity." In Farwell, David, Gerber, Laurie, and Eduard Hovy (Eds.) *Machine Translation and the Information Soup*. Berlin: Springer Verlag. pp. 164–173.

Beuttenmüller, Brigitte. 1997. "Report on the Workshop on Controlled Language Applications (CLAW96)." In *TC-Forum* magazine, volume 1-97, pp. 6–7.

Brockmann, Daniel. 1997. "Controlled Language and Translation Memory Technology: a Perfect Match to Save Translation Cost." In *TC-Forum*. 4-97. December 1997, pp. 10–11. www.tc-forum.org/topictr/tr06cont.htm

Chandioux, John and Annette Grimaila. 1996. "Specialized Machine Translation." Paper presented at the Association for Machine Translation in the Americas (AMTA) conference, Montreal, Quebec, Canada, 2–5 October 1996.

Chervak, Steve, Drury, Colin, and James Ouellette. 1996. "Field Evaluation of Simplified English for Aircraft Workcards." In Proceedings of the 10th FAA-AAM Meeting on Human Factors in Aviation Maintenance and Inspection. Alexandria, Virginia: Jan., 1996. http://galaxyatl.com/hfami/mtng10/drury.htm

CLAW96: The First International Workshop on Controlled Language Applications (CLAW96), Leuven, Belgium: Centre for Computational Linguistics, Katholieke Universiteit Leuven, 26–27 March 1996. www.ccl.kuleuven.ac.be/claw/programme.html

CLAW98: The Second International Workshop on Controlled Language Applications (CLAW98), Pittsburgh, Pennsylvania: Language Technologies Institute, Carnegie Mellon University, 21–22 May 1998. www.lti.cs.cmu.edu/claw98/schedule-onsite.html

Dachelet, Roland. 1994. *Sur la Notion de Sous-langage*. PhD Thesis. Université de Paris 8.

Farrington, Gordon. 1996. "AECMA Simplified English: An Overview of the International Aircraft Maintenance Language." In CLAW96 proceedings, pp. 1–21.

Farrington, Gordon. 1997. "AECMA Simplified English version of the CLAW96 report." In *TC-Forum* magazine, volume 1-97, pp. 10–11.

Goyvaerts, Patrick. 1996. "Controlled English: Curse or Blessing?—A User's Perspective." In CLAW96 proceedings, pp. 137–142.

Grishman, R. and R. Kittredge, eds. 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Hillsdale, New Jersey: Lawrence Erlbaum.

Holmback, Heather, Shubert, Serena, and Jan Spyridakis. 1996. "Issues in Conducting Empirical Evaluations of Controlled Languages." In CLAW96 proceedings, pp. 166–177.

Janssen, Gerd, Mark, Gerhard, and Bernd Dobbert. 1996. "Simplified German—A Practical Approach to Documentation and Translation." In CLAW96 proceedings, pp. 150–158.

Lewis, Terence. 1999. "The Best of Both Worlds—or, Will Two Mongrels Ever Make a Pedigree?" Paper presented at ASLIB-TC21.

Kamprath, Christine, Adolphson, Eric, Mitamura, Teruko, and Eric Nyberg. 1998. "Controlled Language Multilingual Document Production: Experience with Caterpillar Technical English." In CLAW98 proceedings, pp. 51–61. www.lti.cs.cmu.edu/research/kant/pdf/claw98ck.pdf

Knops, Uus and Bart Depoortere. 1998. "Controlled Language and Machine Translation." In CLAW98 proceedings, pp. 42–50.

Mark, Gerhard. 1997. "Didactic-Typographic Visualisation (DTV) version of the CLAW96 report." In *TC-Forum* magazine, volume 1-97, pp. 8–9.

Means, Linda. 1999. Message entitled "re: expanding text" posted on the Controlled Languages List (controlled-languages@let.uu.nl) by linda.means@gm.com, 13 May 1999.

Mellor, Paul. 1998. "The Introduction and Use of Controlled Language at Rolls-Royce." Presented at the annual conference of Technical Information Creation and Distribution (TICAD), 17 November 1999, Birmingham, England.

Mitamura, Teruko. 1999. "Controlled Language for Multilingual Machine Translation." Paper presented at MT Summit VII, Singapore, 13–17 September 1999. www.lti.cs.cmu.edu/research/kant/pdf/mtsummit99.pdf

Mitamura, Teruko and Eric Nyberg. 1995. "Controlled English for Knowledge-Based MT: Experience with the Kant System." Paper presented at 6th International Conference on Theoretical and Methodological Issues (TMI) in Machine Translation. Leuven, Belgium, 5–7 July 1995. www.lti.cs.cmu.edu/research/kant

Nyberg, Eric, Kamprath, Christine, and Teruko Mitamura. 1998. "The Kant Translation System: from R&D to Large-Scale Deployment." In *Lisa Newsletter* 2, no. 1 (March 1998). www.lti.cs.cmu.edu/research/kant/pdf/lisanews.pdf

Ostler, Nicholas. 1999. "The Limits of my Language Mean the Limits of my World: Is Machine Translation a Cultural Threat to Anyone?" Presented at the Theoretical and Methodological Issues in MT conference (TMI 99), Chester, England, 23–25 August 1999.

Schubert, Klaus. 1999. "Resource and Workflow Management Support and Localization." Paper presented at ASLIB-TC21.

Shubert, Serene, Spyridakis, Jan, Holmback, Heather, and M.B. Coney. 1995. "The Comprehensibility of Simplified English in Procedures." In *Journal of Technical Writing and Communication*, 25, no. 4.

SimulTimes. 1999. "Overview of Controlled Language." In *SimulTimes Globalization Newsletter*. 4 October 1999.

Van der Eijk, Pim. 1997. "Controlled Languages in Technical Documentation." In *Computational Linguistics in the Netherlands 1997*, edited by Peter-Arno Coppen, Hans van Halteren, and Lisanne Teunissen and abridged version in *Elsnews* newsletter, February 1998.

Wittgenstein, Ludwig. 1921. *Tractatus Logico-Philosophicus*.