How are we responding to industrial and business needs for Controlled Language and Machine Translation

Journées Linguistiques – Langues contrôlées; traduction automatique et langues spécialisées: 5-6 May 2004
Besançon, France

Jeff ALLEN (jeff.allen@free.fr)

http://www.geocities.com/jeffallenpubs/besanconmay2004.htm

Questions?

- ν Are we responding to the needs of industry?
- ν Are we looking at real business user needs?

What is product lifecycle and life expectancy?

ν What is the current product lifecycle to market of computer equipment?

answer: Approx. 4 months

ν What is the life expectancy of a computer?

answer: Approx 3-5 years

What is end-to-end design to deployment lifecycle?

ν What is the prototype design / development / deployment / implementation timeframe for a new major aircraft (e.g, an Airbus 320 plane)

answer: Approx 7-10 years (yet taking less uptake time with each new aircraft version per comments by Florence B at this conference)

How much volume?

- ν How much technical user and service documentation is published for an aircraft and accompanies that aircraft?

  answer (cited from Barthe, 2004):

  - 50 Manuals (Maintenance, repair, engineering, operations, ..)
  - Large volumes: A320: 120 000 pages + films + aperture cards = 8 T paper
  - Regular updates
  - 2000: 480T doc shipped = 100 M printed pages

What is the CL lifecycle?

ν   What is development / implementation timeframe for monolingual or multilingual CLs in industrial contexts?

answer:

5-10 years for industrial projects
1 month to 2-4 years for research projects

Monolingual CLs – industry & research

- Basic English (1930s)
- Caterpillar Fundamental English (CFE) 1970s
- International Language of Service and Maintenance (ILSAM)
- Bull Global English
- Perkins/Univ Edinburgh PACE
- AECMA Simplified English (SE)
- GIFAS Rationalised French
- Kokak International Service Language
- Smart Controlled English
- General Motors Global English
- Securities and Exchange Commission (SEC) Plain English
- Fight the Fog (European Commission)
- MultiDoc project Controlled Languages
- Remedios Ruiz/Richard Sutcliffe Controlled Spanish

Multilingual CLs – industry

- ν Caterpillar Technical English (CTE)
- ν Attempto Controlled English (ACE)
- ν Alcatel COGRAM
- ν Xerox Multilingual Customized English
- ν Kodak International Service Language
- ν General Motors Controlled Automotive Service Language (CASL)
- ν IBM Easy English
- ν ProLingua LinguaNet
- ν Diebold Controlled English

Multilingual CLs – industry and research

- ν Scania Swedish
- ν Smart Controlled English
- ν Nortel Standard English (NSE)
- ν OCÉ Controlled English
- ν Sun Controlled English
- ν Avaya Controlled English
- ν Oracle ORACAL
- ν Allen's Controlled English for DIPLOMAT project

Controlled Language (CL) commercial off-the-shelf (COTS) products

- ν MAXit checker
- ν Boeing SE Checker
- ν DFKI checker
- ν Lantworks CL checker
- ν PENG (prototype in progress)
- ν what happened to all the others (Clearcheck, Cleaver, etc) ?

Completeness of CL Rule Sets

Taken from O'Brien, 2003

Complete:
ν Aecma SE, CL1, CL4

Reasonably complete:
ν ACE, CL2, CL3, CL6

Incomplete:
ν CL5

Summary: CL rule comparison

Taken from O'Brien, 2003

- ν Only one CL rule common to all eight CLs

- ν Only 7 CL rules shared by at least four of the CLs

- ν Half of the CLs have over 50% unique rules

- ν CLs in this analysis do not have a lot in common and there is no obvious "core" set of Controlled English rules…

Explanation of CL rule differences

Taken from O'Brien, 2003

Differences caused by:

ν Objectives of rule sets

ν MT system and language direction

ν Influence from corporate writing rules/style guides

ν Sheer subjectivity

ν Acceptability for Authors

Continuum Analysis

Results from another analysis of CL rules along the continuum of human vs machine-oriented CLs (from Reuther, 2003)

- ν Comprehensability relevance (Readability: R-rules)
- ν Translational relevance (Translation: T-rules)

Comparison of R- and T- rules

Taken from Reuther, 2003

- ν all R-rules (42) are included in the set of T-rules (59)

- ν the divergent rules (17) are considered relevant for translation but not for readability

- ν 10 with low priority

- ν 7 with high priority

Conclusion of R- vs T-rule analysis

Taken from Reuther, 2003

- ν Readability requires less rules than translatability

- ν R- rules are subset of T-rules

- ν the two rule sets differ only slightly

- ν T-rules are more "specific" than R-rules

- ν both R-rules and T-rules as part of a CL depend on the intended user scenario

Monolingual vs. Multilingual

- ν Monolingual-Oriented Controlled Language
  - Only for a single language or also possible to create CLs for several languages based on a core set of rules

- ν Multilingual-Oriented Controlled Language
  - via multilingual (primarily machine-oriented yet also human) translation processing techniques

Translation Approach – Inbound vs. Outbound

ν   Inbound Translation

(aka content gisting, assimilation, etc)

ν   Outbound Translation approach

(aka translation for publication, dissemination, etc)

Determine your need, your approach, and then select the tool that corresponds to your need

Machine Translation (MT) types

- ν Rule-Based MT (RBMT)
- ν Example-Based MT (EBMT)

  (Translation Memory)
- ν Knowledge-Based MT (KBMT)
- ν Statistical MT (SMT)
- ν Multi-Engine MT (MEMT)

Machine Translation (MT) commercial off-the-shelf (COTS) products

ν   Promt-based products

   (Promt, Reverso, e-promt)

ν   Systran products

ν   SDL Transcend

ν   Translation Experts InterTran

ν   Linguatec

ν   Comprendium

ν   iTranslator (Mendez/Bowne Global)

ν   Lantworks (was METAL) (Lant//Xplanation)

ν   CIMOS system

ν   SAKHR software

Conclusions

- ν Creating a corporate-specific CL with integrated translation technologies often takes about 5-6 years to implement (there are exceptions!)
- ν Projects are re-inventing the wheel each time
- ν Need to focus more on naming and approaches such as "International writing and technical communication guidelines" and "translation software" for these to be more psychologically acceptable to users than the terms "controlled language" and "machine translation"

Conclusions

- ν Very few examples (other than AECMA SE) of writing in CL s are publicly available. We need sharable CL and MT corpora
- ν See publications on CLs via my portal site (www.geocities.com/jeffallenpubs)

References

- Allen, Jeff. Controlled Translation issues : implementing Controlled Language and Machine Translation. Presented at EAMT/CLAW2003, Dublin, Ireland, 15-17 May 2003. http://www.ctts.dcu.ie/presentations.html

- Barthe, Kathy. 2004. Airbus documentation and Controlled English. Invited talk presented at the France Chapter of the Society for Technical Communication. 2 April 2004

- O'Brian, Sharon. An Analysis of Several Controlled English Rule Sets.   Presented at EAMT/CLAW2003, Dublin, Ireland, 15-17 May 2003 http://www.ctts.dcu.ie/presentations.html

References

- Reuther, Ursula. Two in One – Can it Work? Readability and Translatability by means of Controlled Language. Presented at EAMT/CLAW2003, Dublin, Ireland, 15-17 May 2003. http://www.ctts.dcu.ie/presentations.html
- KANT project: http://www.lti.cs.cmu.edu/Research/Kant/