Controlled language for MT in action

- Johann Roturier
  - Symantec Ireland

- Translingual Europe, 14 May 2009

Outline

1 Overview of CL

2 CL deployment

3 Lessons learnt

4 Impact of CL on MT

5 Future directions

Overview of Controlled Language

- Strict use of Controlled Language (CL)
  - “Subset of a natural language that uses a restricted grammar and a restricted vocabulary” (technical domain)
  - Makes source content clearer and less ambiguous
  - Improves comprehensibility and (machine-)translatability of source content
  - Example: Caterpillar Technical English in the mid 1990s (142 rules and more than 70000 terms)

- Loose use of Controlled Language
  - Used since 2006 for large structured documentation sets (authored in XMetal)
  - MT requirements in FR, DE, IT, BR, ES, ZH and JA (SYSTRAN)
  - Strict enforcement of spelling and grammar checks
  - Enforcement of corporate terminology once defined
  - Enforcement of specific CL rules
    - Minimize impact on authoring productivity

CL deployment

- Use authoring application to implement CL rules for English
  - Language checker developed by acrolinx
  - Customized to deal with Symantec content
  - Based on pattern-matching approach (reactive approach)
  - Rules can be context-sensitive using XML documents mark-up (DocBook)
  - Suggestions or help files are provided to users

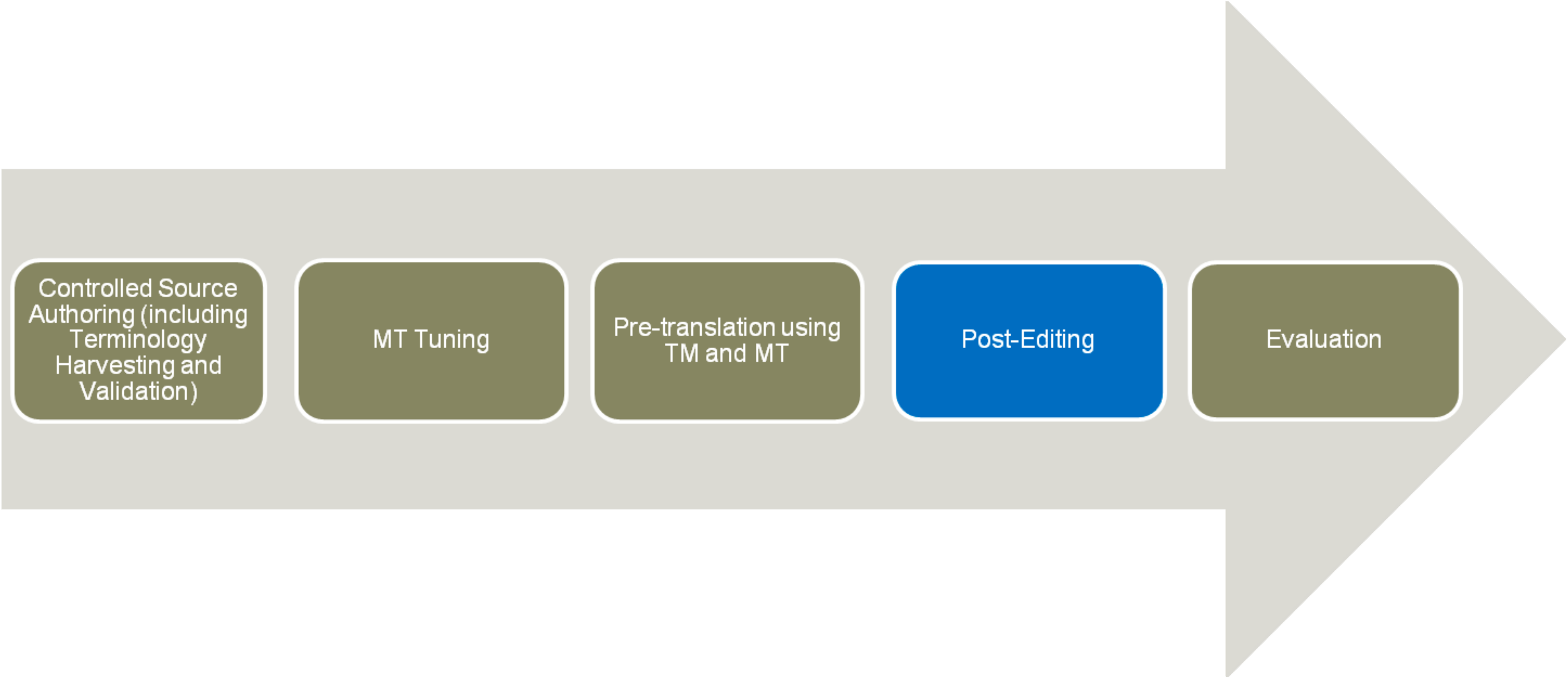
- acrolinx IQ™ is used to ensure that source content is compliant with
  - Grammar rules
  - Terminology (5000+ terms)
  - Rules based on corporate guidelines (20+)
    - Some rules deal specifically with tagging, such as the tagging of SW references
  - MT-specific rules (6)

- acrolinx IQ™ is used to harvest, store and access source terminology
  - In-house panel working to further refine rule set and terminology lists

Lessons learnt during CL deployment

- Strict implementation when there is:
  - New content
  - Little leverage
  - Time
- Resources must be maintained
  - Eliminate false alarms
  - Adapt resources to deal with new content
  - Language-specific rules are best implemented as:
    - Pre-processing step
    - MT Normalisation dictionaries
- CL + MT is not always sufficient
  - Terminology work to update dictionaries (15000+ entries per dictionary)
  - PE when specific qualify standard is required

Global content delivery workflow using CL and MT



Controlled Source Authoring (including Terminology Harvesting and Validation)

MT Tuning

Pre-translation using TM and MT

Post-Editing

Evaluation

Importance of CL for (RB)MT

- Why is it so difficult to use MT effectively?
  - While attempting to install SystemWorks 2005 the error message "Error 1722. An error occurred while performing the task. There is a problem with this Windows Installer package. A program run as part of the setup did not finish as expected. Contact your support personnel or package vendor" appears.

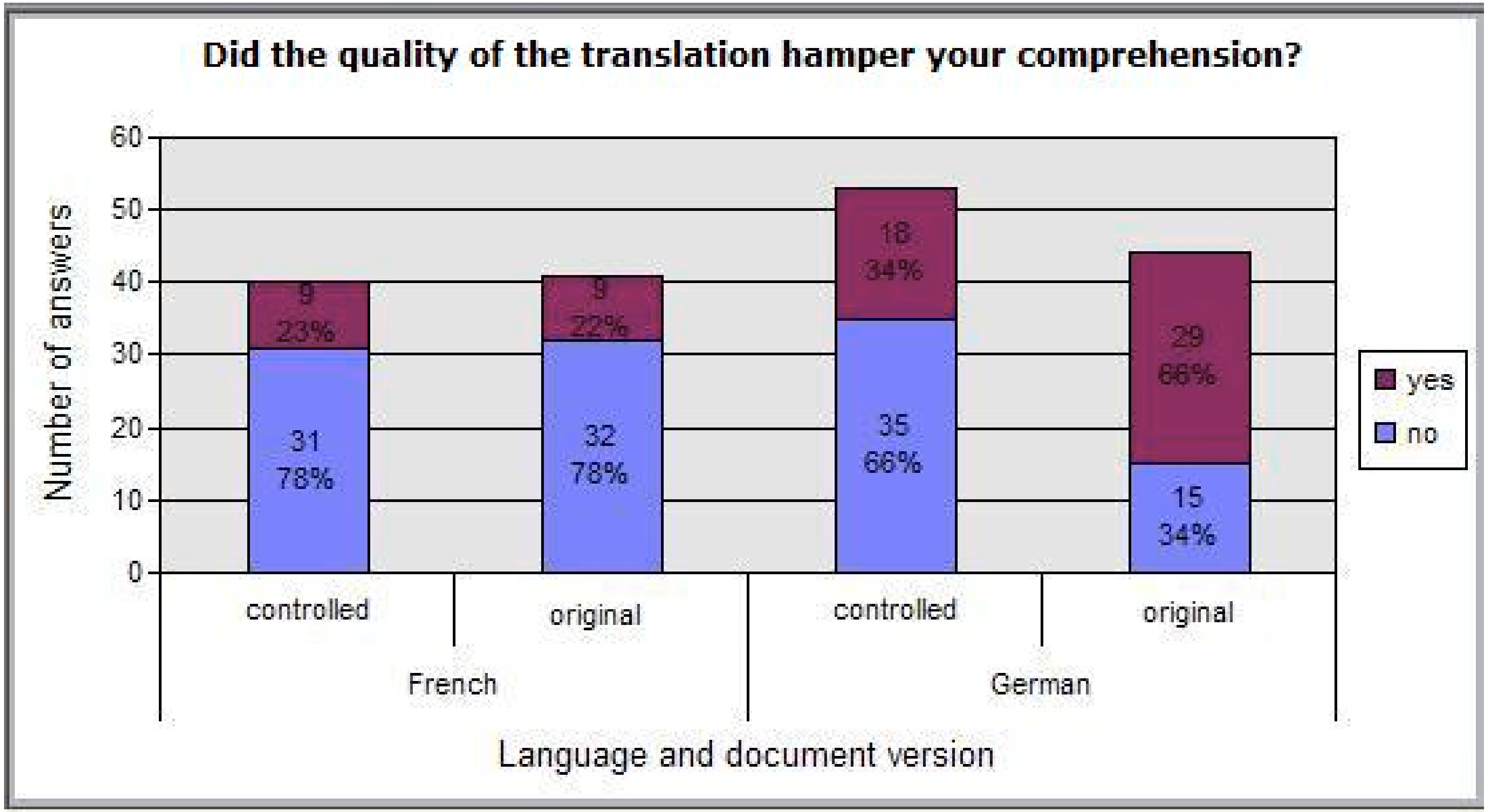
- Controlling source content at the segment level
  - Reduces complexity and ambiguity during MT step
  - Reduces post-editing effort during PE step

- Controlling source content at the sub-segment level (terminology)
  - New terms harvested and defined during authoring
    - Identified variants can be used by search engine (for help system)
    - Reduces MT tuning step
  - Approved translations are defined during MT tuning step

Impact of source compliance on MT quality

Source words	MT quality	Evaluation type	acrocheck™ project score
1083	Excellent	Human	28
3677	Good	Human	79
2546	Medium	Human	118
2129	Poor	Human	150
10972	Greater than 0.6 GTM scores	Automatic	64
9926	Less than 0.6 GTM scores	Automatic	147

Impact of source compliance on comprehensibility



Future directions

- Use automatic tagging and pre-processing
  - Tagging: New tags may be used to mark ambiguous terms and used to produce context-sensitive translations (Systran XSL)
  - Pre-processing: add, remove or move source words or phrases

- Combine a CL approach with a semantic reuse approach
  - Increase 100% TM matches by leveraging TMs during authoring (CL through example)

- Use CL approach to check MT output
  - Can repetitive MT errors be flagged? And possibly fixed automatically?

- Use CL knowledge to deal with uncontrolled content
  - Short technical notes cannot always go through a full CL cycle, community generated content is increasing (e.g. forums, chat)
  - Term variants and deprecated terms can be added to MT normalisation dictionary