Outline

- Introduction
  - What is Controlled Language?
  - Goals of Controlled Language
  - Types of Controlled Language
  - Advantages and Challenges
- History of CL & Applications
  - Document Authoring
  - Document Translation

Outline [2]

- Designing a Controlled Vocabulary and Grammar
- Deployment Issues for CL
- Evaluating the Use of Controlled Language
- Automatic Rewriting for MT

Introduction

What is Controlled Language?

- A form of language usage restricted by grammar and vocabulary rules
- No single “controlled language” for English
- Controlled language can be used:
  - solely as a guideline for authoring
  - with a checking tool to verify conformance
  - in conjunction with machine translation

Goals of Controlled Language

- Achieve consistent authoring
- Encourage clear and direct writing
- Improve the quality of translation output
- Use as input to machine translation systems
e.g. The KANT System, CASL System
Types of Controlled Language

- **Human-oriented CL**: to improve text comprehension by humans (for authors and translators)
- **Machine-oriented CL**: to improve “text comprehension” by computers (for CL checkers or MT systems)

Designing for Different Types of CL

Examples of Writing Rules

- Do not use sentences with more than 20 words
- Do not use passive voice
- Do not make noun clusters of more than 4 nouns
- Write only one instruction per sentence

Examples [2]

- Make your instructions as specific as possible
- Use a bulleted layout for long lists
- Present new and complex information slowly and carefully

Q: Which rules can be checked automatically?

CL Advantages

- Improves the source text:
  - readability
  - comprehensibility
  - consistency
  - reusability
- Improves translation:
  - controlled texts easier to translate
  - consistent text easier to reuse

CL Challenges

- Writing may become more time-consuming
- An additional verification step is required
- Developing a CL may be costly
- CL use must be evaluated carefully
History of CL & Applications

Roots of CL

- C.K. Ogden’s “Basic English” (1930’s)
  - 850 basic words
  - an “international language”, foundation for learning standard English
  - never widely used

Roots of CL [2]

- Caterpillar Fundamental English (CFE) - 1970’s
  - Non-technical vocabulary and grammar
  - First version had only 850 terms
  - For non-native English speakers
  - Abandoned after ~10 years:
    - insufficient for complex writing
    - CFE difficult to train and enforce

Examples

Non CFE: “Enlarge the hole.”
CFE: “Use a drill to make the hole larger.”

Non CFE: “The brake components must be matched during installation.”
CFE: “The brake parts with same numbers on the lower ends of the brake shoes must be installed together.”

Survey of CLs

- Ogden’s Basic English
- Caterpillar Fundamental English (CFE)
- Smart’s Plain English Program (PEP)
- White’s International Language for Serving and Maintenance (ILSAM)
- AECMA
- IBM
- Ericsson Telecom
- Rockwell International
- Hyster
- Boeing SE

CL Checking

- Aids an author in determining whether a text conforms to a particular CL
  - Verify all words & phrases are approved
  - Verify all writing rules are obeyed
  - May offer help to the author when words or sentences not in the CL are found
CL for Machine Translation
• Use of software to analyze texts and translate to other languages
• Technical Translation
  – Large segment of translation market
  – Documentation for complex products (e.g., consumer electronics, computer hardware, heavy machinery, automobiles, etc.)
  – Involves large, specialized vocabulary
  – Writing style may be complicated

Challenges for MT
• Ambiguity
  – Lexical, Structural, Referential
• Complexity
  – Assigning meaning to complex syntactic structures
• Controlled language reduces the impact of these phenomena while increasing source text quality

Designing a Controlled Vocabulary and Grammar

Controlled Vocabulary
• Restrict vocabulary size and meaning
• Most useful way to limit ambiguity of input sentences
• Key to improve the accuracy of translation

Encoding the Meanings of Vocabulary Items
• Limit Meaning per Word/Part of Speech Pair
  – Helps to reduce the amount of ambiguity
• Encode Meanings Using Synonyms
  – Finding separate, synonymous terms
  – Encode them in the lexicon
  – Synonymous terms are marked in the lexicon
  – Used in support of on-line vocabulary checking

Encode Truly Ambiguous Terms
• When a term must carry more than one meaning in the domain
• Encode in separate lexical entries
• Resulting output structure will be ambiguous
• Lexical disambiguation by machine or by author
Designing a Controlled Grammar

- What is CL used for?
  - Authoring without CL checker?
  - Authoring with CL checker?
  - Translating with MT?
  - Translating without MT?
- What types of constraints are needed?
- Design focus: to reduce ambiguity

Problematic Structures

- Use of participial forms (such as -ing and -ed)
  - Used in a subordinate clause without a subject
    “When starting the engine…”
  - Reduced relative clauses
    “the pumps mounted to the pump drive”

Problematic Structures [2]

- Verb Particles “turn on” → “start”
- Coordination of Verb Phrases
  “extend and retract the cylinders”
- Conjoined Prepositional Phrases
  “pieces of glass and metal”
- Quantifiers and Partitives
  “repeat these steps until none are left”

Problematic Structures [3]

- Coordinate Conjunction of S
  (conjuncts must be the same type)
- Adjoined Elliptical Modifiers
  “if necessary”, “if possible”, “as shown”, etc.
- Punctuation - rules for consistency
  - use of comma, colon, semi-colon
  - quotation marks
  - parentheses

Problematic Structures [4]

- Relative Clauses - should be introduced by relative pronouns
- Subject gap relative clause
  “The service man can determine the parts which are at fault”
- Object gap relative clause
  “The parts which the service man orders”

Deployment Issues for CL
Deployment Issues for CL

• CL cannot be too strict
• Author usability and productivity are important for deployment
• Expressiveness -- Balance vocabulary size vs. complex grammatical expressions
• Productivity of authoring vs. Post-editing

Deployment Issues for CL (2)

• Controlled Target Language Definition
  – Translated documents at the same stylistic quality level as the source documents
  – Set appropriate expectations about translation quality
  – Controlled language specification for TL
  – Produces more useful aligned corpora for TM

Deployment Issues for CL (3)

• Controlled Language Maintenance
  – Need to update the terminology and grammar
  – Requires a well-defined process that includes the customer/user:
    • Problem reporting
    • Initial screening of the problems
    • Process monitoring and quality control
    • Support rapid terminology and grammar updates for source and target languages

Success Criteria for CL

• Translation for Dissemination
• Highly-Trained Authors
• Use of Controlled Language Checker
• Technical Domain

Benefits of CL

• Improved consistency of writing
• Increased re-use of documents
• Improved authoring quality
  – value of writing guidelines, term management
  – value of standardized authoring
  – improved quality / consistency of training

Evaluating the Use of Controlled Language
Benefits of CL

- Useful for reducing ambiguity
- Ambiguity Test:
  - Average # of syntactic analyses per sentence dropped from 27.0 to 1.04
  - 95.6% have a single meaning representation
  - Lexical constraints achieve the largest reduction in ambiguity
- Improve the quality of translation output

Challenges

- Domain ambiguity is pervasive
- Terminology maintenance can be costly
- For writers and translators, style is more satisfying than productivity, consistency, simplicity, ...
- For end users, simplicity and clarity are a top priority

CL in the Real World

- Software performance (shouldn’t impact on author productivity)
- Author commitment (writing well vs. “getting it to pass”)
- Organizational commitment (publishing deadlines vs. CL compliance)

Specification vs. Coverage

CL is Justified When ...

- Benefits a large document volume
- Documents are hierarchical, reusable
- Checking well-integrated with document production system
- Controlled source reduces cost of translation to multiple target languages
Recent CL Developments

- CL for Technical Documentation
  - AECMA’s Simplified English (SE)
  - Caterpillar Technical English (CTE) by KANT
  - Boeing Simplified English Checker (BSEC)
  - GM’s Controlled Automotive Service Language (CASL)
  - Easy English (IBM)

KANT Controlled Language Checker

- Thin-client checker program runs on author’s PC (Java)
- Accesses KANT analyzer software running on a network server
- Features:
  - dynamic checking (while the author is typing)
  - automatic PP disambiguation
  - pronoun resolution
  - grammar diagnostics

Analysis of CL Rewriting

- Studied author logs from sessions with the authoring tool (heavy equipment domain)
- The log files contained 180,402 sentences
- 94% of the sentences did not require rewriting
- For 1461 sentences (0.8%) the author attempted 4 or more rewrites

<table>
<thead>
<tr>
<th># of Attempts for Rewriting</th>
<th># of Rewrites</th>
<th>Total Sentences</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>169,505</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5,404</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2,792</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,240</td>
<td>0.7%</td>
</tr>
<tr>
<td></td>
<td>4 - 45</td>
<td>1,461</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>180,402</td>
<td>100%</td>
</tr>
</tbody>
</table>

Analysis of CL Rewriting (2)

- We also analyzed sentences from a different domain (laser printer manual)
- Identified constructions which have the greatest impact on author productivity
- Found most common problems

Most Common Problems

- Unknown Noun Phrase
  - KANT Controlled English (KCE) does not allow arbitrary noun-noun compounding
- Missing Determiner
- Coordination of Verb Phrases
- Missing or Improper Use of Punctuation
Most Common Problems (2)

- Missing “in order to” phrase
  - In KCE, purpose infinitival clause should use “in order to” instead of “to”
- Use of “-ing” form
  - In KCE, “-ing” cannot be used immediately after a noun (e.g. The engine sends the information indicating that ...)

Most Common Problems (3)

- Coordination of Adjective Phrases
  - In KCE, adjective coordination before a noun is not allowed
    - Non-KCE: top left and right sides
    - KCE: the top left side and the top right side
- Missing Complementizer “that”
  - “that” cannot be omitted in KCE
  - Non-KCE: Ensure it is set properly

Grammar Diagnostics

- 2 New modules, Diagnostifier (full syntactic analysis) and PatternFinder (pattern matching), were added to the KANTOO architecture
- Diagnostifier and PatternFinder determine whether or not a particular sentence triggered certain diagnostic rules in the CL grammar
- If so, a detailed message is prepared
- The message is transmitted to the CL Checker
- A specific user dialog is invoked

Design of Grammar Diagnostics

- If the set of possible parses for an input contains at least one f-structure without diagnostics, then the parse continues to the Disambiguation Module.
- If all f-structures contain diagnostics, they are passed to the Diagnostifier.
  - Scores of all diagnostics within each f-structure are summed.
  - The f-structure with the lowest total score is preferred. In case of a tie, the system picks one arbitrarily.
  - The relative scores associated with diagnostics were determined by trial and error.
  - If the best f-structure (lowest total score) has more than one diagnostic, the diagnostic with the lowest score is presented to the user first.

Scores for Diagnostics

<table>
<thead>
<tr>
<th>Diagnostics</th>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MISSING_DET</td>
<td>Determiner missing before noun</td>
<td>10</td>
</tr>
<tr>
<td>UNKNOWN_NP</td>
<td>Noun phrase not in the dictionary</td>
<td>10**</td>
</tr>
<tr>
<td>IN_ORDER_TO</td>
<td>Missing “in order to”</td>
<td>12</td>
</tr>
<tr>
<td>MISSING_PUNC</td>
<td>No period at the end of sentence</td>
<td>13</td>
</tr>
<tr>
<td>BY_USING</td>
<td>Need “by” before “using”</td>
<td>15</td>
</tr>
<tr>
<td>VP_COORD</td>
<td>Two verbs cannot be conjointed</td>
<td>15</td>
</tr>
<tr>
<td>MISSING_THAT</td>
<td>Use complementizer “that”</td>
<td>15</td>
</tr>
<tr>
<td>ADJ_COORD</td>
<td>Two adj. cannot be conjointed</td>
<td>16</td>
</tr>
<tr>
<td>IMPROPER_PUNC</td>
<td>Do not end noun phrase in a period</td>
<td>21</td>
</tr>
<tr>
<td>IMPROPER_ING</td>
<td>Bad use of an “-ing” form</td>
<td>25</td>
</tr>
</tbody>
</table>

* 10 for phrases, else 11; ** 10 if standalone, else 20
Diagnostic Algorithm (2)

- If the sentence doesn’t parse, PatternFinder tries to find a problem.
- If PatternFinder can’t find a problem, the Parser returns the general message: “The sentence is not grammatical.”

Two Types of Diagnostics Using KANTOO Syntactic Parser

1. Offer a diagnostic message and rewrite for a sentence
   - Missing Determiner
   - Missing Complementizer “that”
   - Missing or Improper Use of Punctuation
   - Missing “in order to” phrase
   - Missing comma
   - Etc.

Diagnostic Message: Interactive Rewriting

Click on the button to receive the channel settings.

Diagnostic Message: Unknown NP

“all filter change”

“all filter change” is not in the dictionary.

Two Types of Diagnostics Using KANTOO Syntactic Parser (2)

2. Offer a diagnostic message only
   - Unknown Noun Phrase: a lexicographer needs to decide whether to add the term to the lexicon
   - When –Ving: what is the subject of the clause? Usually, it is the same as main clause subject, but not always.

Diagnostics by Pattern Matching

1. With a message and rewrite
   - Contraction: e.g. “you’re” “haven’t”
   - “have to”: change to “must”
   - “whether or not”: change to “whether”
   - etc.

2. With a message only
   - Quotes, semicolon, dash, reflexive, etc.
Evaluation (1)

- 4229 non-KCE sentences were tested from computer printer manuals.
- 2843 sentences (67.2%) received a diagnostic message.
  - 1741 sentences (60%) exhibited grammar diagnostics.
  - 1129 sentences (40%) exhibited a diagnostic of unknown single terms.


Results from Randomly-selected Documents

<table>
<thead>
<tr>
<th>Diagnostics</th>
<th>No. Sentences</th>
<th>No. Correct</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown Term</td>
<td>234</td>
<td>234</td>
<td>100%</td>
</tr>
<tr>
<td>Grammar</td>
<td>603</td>
<td>521</td>
<td>86.4%</td>
</tr>
<tr>
<td>Total</td>
<td>837</td>
<td>755</td>
<td>90.2%</td>
</tr>
</tbody>
</table>

Results of Automatic Rewrites

<table>
<thead>
<tr>
<th>Grammar Diagnostics</th>
<th>No. Sentences</th>
<th>No. Correct Rewrites</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offer Rewrites</td>
<td>312</td>
<td>279</td>
<td>89.4%</td>
</tr>
</tbody>
</table>

Evaluation (2)

- 1302 sentences were tested, in which authors tried to rewrite 4 or more times before passing KCE.
- 569 sentences (44%) received a diagnostic message.
  - 415 sentences (32%) exhibited grammar diagnostics.
  - 154 sentences (12%) exhibited a diagnostic of unknown single terms.
- 733 sentences (56%) did not receive a diagnostic message.
  - Most of the problems were from obsolete SGML tagging.
  - Other problems: Incomplete sentences, comparative, etc.


Results

<table>
<thead>
<tr>
<th>Diagnostic</th>
<th>No. Sentences</th>
<th>No. Errors</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>MISSING NP</td>
<td>240</td>
<td>12</td>
<td>95%</td>
</tr>
<tr>
<td>UNKNOWN TERM</td>
<td>154</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>MISSING_DET</td>
<td>60</td>
<td>14</td>
<td>76.6%</td>
</tr>
<tr>
<td>VP_COORD</td>
<td>32</td>
<td>1</td>
<td>96.8%</td>
</tr>
<tr>
<td>MISSING_PUNC</td>
<td>27</td>
<td>2</td>
<td>92.5%</td>
</tr>
<tr>
<td>IMPROPER_PUNC</td>
<td>25</td>
<td>4</td>
<td>84%</td>
</tr>
<tr>
<td>IN_ORDER_TO</td>
<td>15</td>
<td>1</td>
<td>93.3%</td>
</tr>
<tr>
<td>IMPROPER_ING</td>
<td>12</td>
<td>1</td>
<td>91.6%</td>
</tr>
<tr>
<td>ADJ_COORD</td>
<td>3</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>MISSING_THAT</td>
<td>1</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>569</td>
<td>35</td>
<td>93.8%</td>
</tr>
</tbody>
</table>

Discussion

- Missing determiners were the most difficult diagnostics.
  - XML tags are required instead of determiners.
  - Some idiomatic expressions.
Next Steps

• Author Productivity: Measure impact of diagnostics on the authors
• Testing of Recall: Determine if there are additional sentences in the test set for which the system should have raised diagnostics, but did not.
• Automatic Rewriting System