



Controlled Language Input/Output

11-731 Machine Translation

Teruko Mitamura

Language Technologies Institute
Carnegie Mellon University



Carnegie Mellon
School of Computer Science

11-731 Machine Translation

1

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



Carnegie Mellon
School of Computer Science

11-731 Machine Translation

2

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Outline [2]

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



Carnegie Mellon
School of Computer Science

11-731 Machine Translation

3

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Introduction



Carnegie Mellon
School of Computer Science

11-731 Machine Translation

4

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



Carnegie Mellon
School of Computer Science

11-731 Machine Translation

5

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



Carnegie Mellon
School of Computer Science

11-731 Machine Translation

6

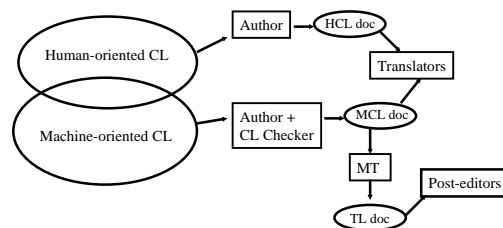
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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."

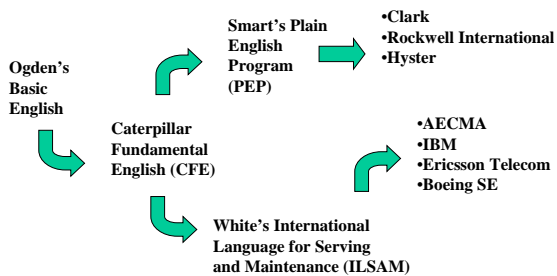
CFE: "Use a drill to make the hole larger."

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



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



Evaluating the Use of Controlled Language



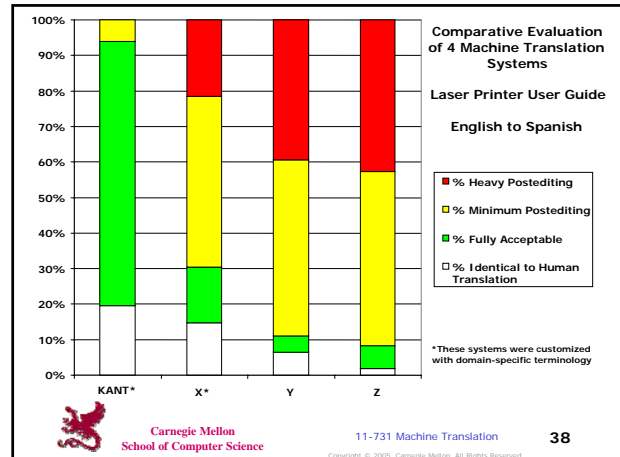
Benefits of CL

- 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



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

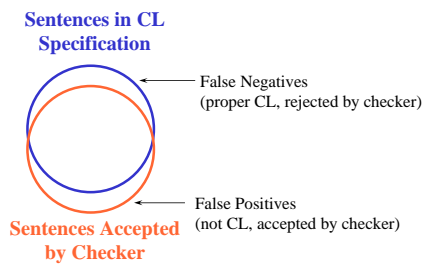


CL in the Real World

- 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



of Attempts for Rewriting

# of Rewrites	Total Sentences	Percentage
0	169,505	94%
1	5,404	3%
2	2,792	1.5%
3	1,240	0.7%
4 - 45	1,461	0.8%
Total	180,402	100%



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*

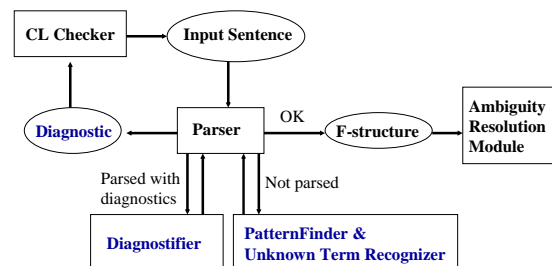


Grammar Diagnostics

- 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



Diagnostic Algorithm

- 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

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

* 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."



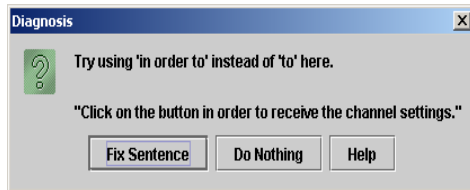
Two Types of Diagnostics Using KANTOO Syntactic Parser

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.

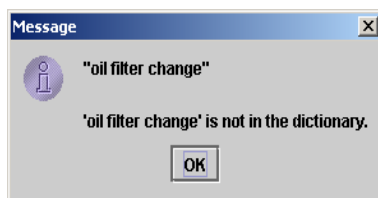


Two Types of Diagnostics Using KANTOO Syntactic Parser (2)

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.



Diagnostic Message: Unknown NP



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

Source: Mitamura, et al. (2003) "Source Language Diagnostics for MT" in Proceedings of MT Summit IX.



Results from Randomly-selected Documents

Diagnostics	No. Sentences	No. Correct	% Correct
Unknown Term	234	234	100%
Grammar	603	521	86.4%
Total	837	755	90.2%



Results of Automatic Rewrites

Grammar Diagnostics	No. Sentences	No. Correct Rewrites	% Correct
Offer Rewrites	312	279	89.4%



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.
- Source: Mitamura, et al. (2003) "Diagnostics for Interactive Controlled Language Checking" in Proceedings of EAMT/CLAW 2003.



Results

Diagnostic	No. Sentences	No. Errors	% Correct
MISSING_NP	240	12	95%
UNKNOWN_TERM	154	0	100%
MISSING_DET	60	14	76.6%
VP_COORD	32	1	96.8%
MISSING_PUNC	27	2	92.5%
IMPROPER_PUNC	25	4	84%
IN_ORDER_TO	15	1	93.3%
IMPROPER_ING	12	1	91.6%
ADJ_COORD	3	0	100%
MISSING_THAT	1	0	100%
Total	569	35	93.8%



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

