Controlled Language and the Implementation of Machine Translation for Technical Documentation

Laura Ramírez Polo

UNIVERSITÄT
DES
SAARLANDES

Contents

1. Motivation and Goal
2. Background: Controlled German and CL Checkers: MULTILINT
3. Evaluating CL Checkers
4. Method Outline
5. Selection of resources
6. Conclusions and Outlook

Motivation and Goal

Evaluation of the Controlled Language Checker MULTILINT

↓

Goal

Develop a method to assess the effectiveness of the implementation of a Controlled Language Checker

Background

- Efforts to establish guidelines for writing technical documentation have resulted in the development of Controlled Languages (CL)

- Their implementation has been frequent in industrial contexts for the past decade

Background

- Benefits of CL:
  - Improvement of Readability and Comprehensibility
  - Improvement of Translatability (human and machine)
- Problems:
  - Difficult to make general statements (for all languages, for all contexts)
  - Lack of standard methods for evaluation

Controlled German and CL Checkers: MULTILINT

Projects MULTILINT and TETRIS (1995-2002):

- Main Partners: IAI, BMW AG
- Goal: “Development of an intelligent linguistic system for the production and administration of technical documentation” (Haller, 01)

Controlled German and CL Checkers: MULTILINT

- MULTILINT aims at controlling the language by helping the authors to write according to a definite set of rules
  - Spelling
  - Grammar
  - Style
  - Vocabulary
  - Terminology

Evaluating CL Checkers

- What should be tested and how it is to be tested (interaction of modules, precision and recall, noise, etc) depends on the context

- Results of tests do not always correlate with effectiveness of CL

Evaluating MULTILINT
TETRIS Project Documentation

- Scenario 1: Human Proof-reading vs. MULTILINT
  - Measurement of Precision and Recall
  - Results: MULTILINT not developed enough to fully substitute human proof-reading
- Scenario 2: Hit Rate in Translation Memory Systems
  - Measurement of increasement of hit rate
  - Results: lack of statistical value, subjective factors

New Evaluation Scenario

- Effectiveness of MTranslatability
  - Evaluate MULTILINT by evaluating the quality of machine translated texts
    - Source text checked with MULTILINT
    - Source text not-checked with MULTILINT

- Context Evaluation: Use of the CL Checker MULTILINT in an industrial context.

Method Outline

1. Selection of resources
    1. Selection of the most suitable text type
    2. Selection of the most suitable MT system
2. Evaluation
    1. Analysis of MULTILINT translatability features for MT
    2. Assessment of effectiveness of MULTILINT´s implementation

The FEMTI-Framework

- Developed within the ISLE-Project (International Standards for Language Engineering)
- Framework for the design of evaluations of MT systems
- Based on the principles of context-based evaluation (Arnold et al. 94)
- Divided in two parts:
    - Evaluation Requirements
    - System characteristics
- Presents evaluation features and different metrics, but proposes no standard metrics

Selection of resources
Context definition

- Industrial environment: e.g. Automotive company
- MULTILINT is applied for the production of technical documentation
- Source language: German
- Target languages: English and probably other languages
- Study MT as a complementary solution to human translation
- Translation task: dissemination (internal and external publication)
- Users: internal users with atomotive background

↓

Declarative Evaluation

Selection of resources
Text type

- Some types of texts are more suitable for MT than others
- Technical documents from automobile domain (repair instructions, training documentation, owner's manuals…) were analysed
- Requirements:
  - Text length
  - Security aspects
  - Compliance with CL (Translatability indicators)
- Results: Selection of repair instructions

Selection of resources
Text corpus

- Text corpus with real texts, 3000 segments for automatic evaluation
- Reduced text-corpus with 250 selected segments for human evaluation, containing:
  - Questionnaire
  - 125 segments for comprehensibility
  - 125 segments for post-editability
  - Final questionnaire

Selection of resources
MT system

- Pre-selection of 3 commercial systems according to
- following criteria:

- Internal characteristics
  - Translation model: rule-based systems
  - Language pairs (Languages)I
  - Terminology (Dictionaries)
  - Status of Vendor
  - Previous evaluation studies
- External characteristics
  - Evaluation with adjustment
  - Output Quality
    - Comprehensibility and Post-Editability (Human evaluation)
    - Fidelity through BLEU (as proposed by FEMTI)

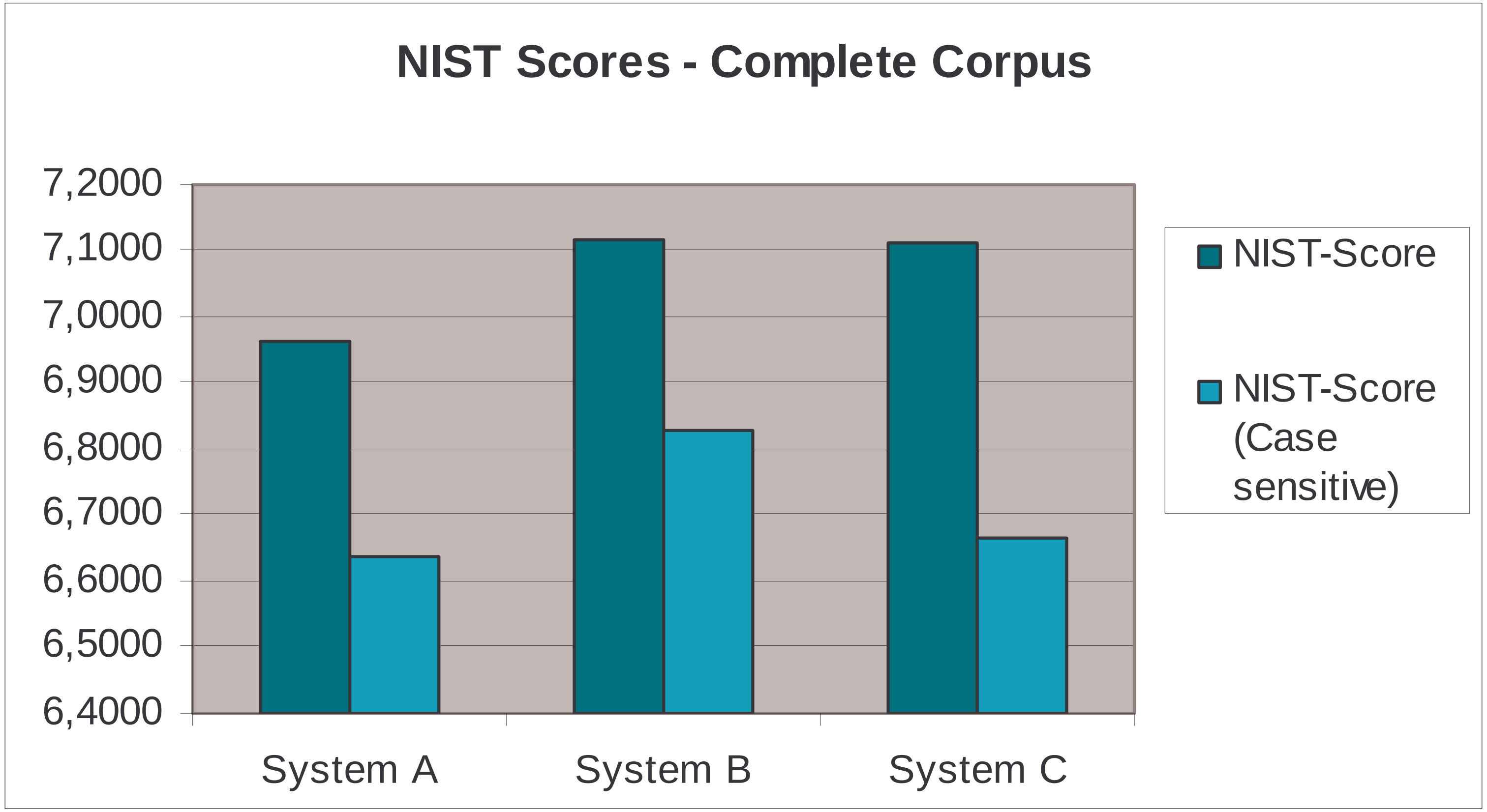
Output Quality: Evaluation Metrics

- Automatic Metrics
  - n-gram based metrics (BLEU, NIST)
  - Advantages: cost-effective, objective, reproducibility and comparability
  - Pitfalls: not always reliable, callibration with human results required, interpretation not clear, only for evaluating homogeneous systems
- Human Metrics
  - Scales, Questionnaires
  - Advantages: results pretty significant
  - Pitfalls: costly, time-consuming, hardly reusable, subjective

Automatic Evaluation

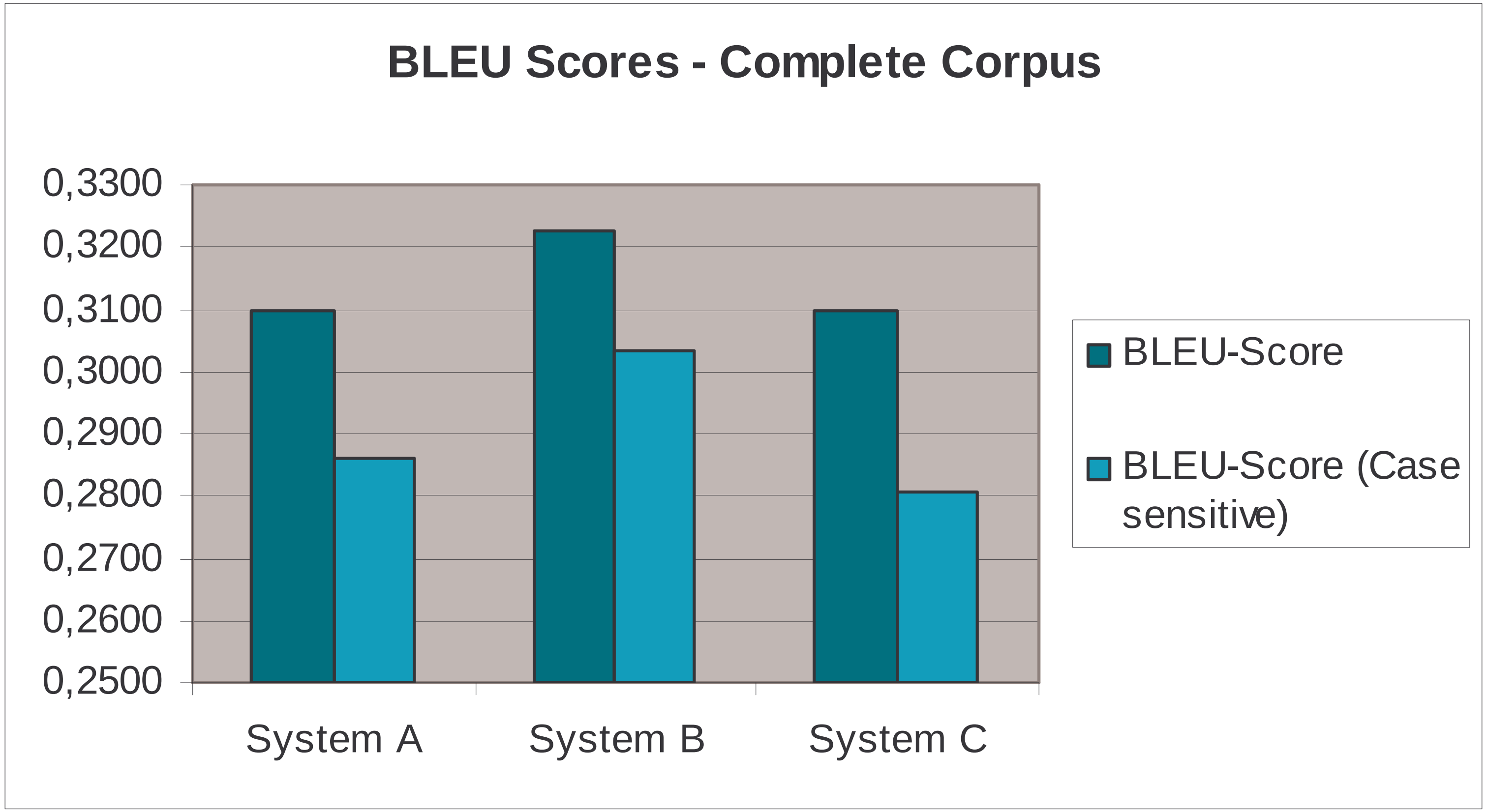
- MT evaluation kit (NIST)
- BLEU and NIST metrics
- Evaluation of whole and reduced corpora
- Only one human reference translation (free human translation)

NIST Results
Complete Corpus



UNIVERSITÄT DES SAARLANDES

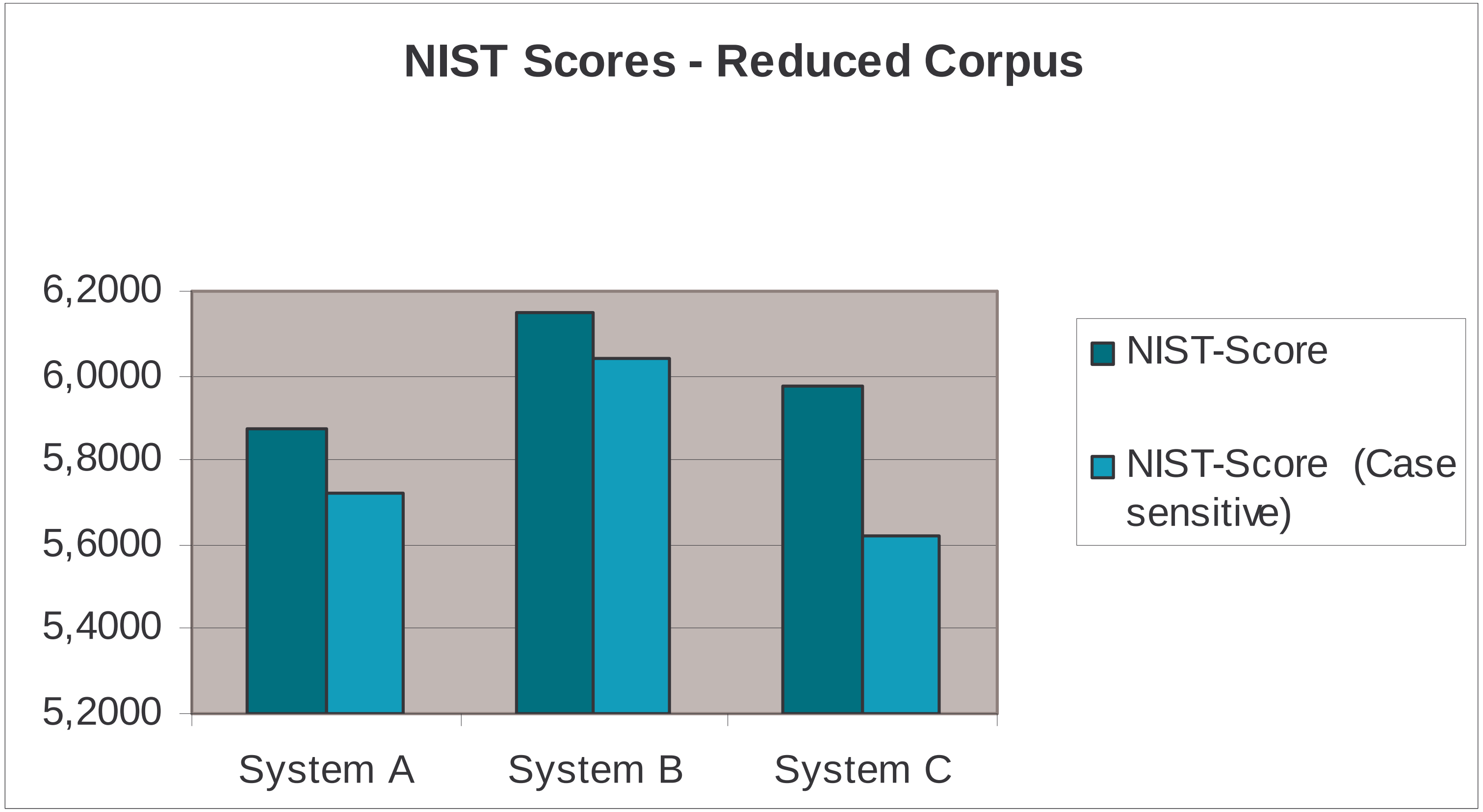
BLEU Results
Complete Corpus



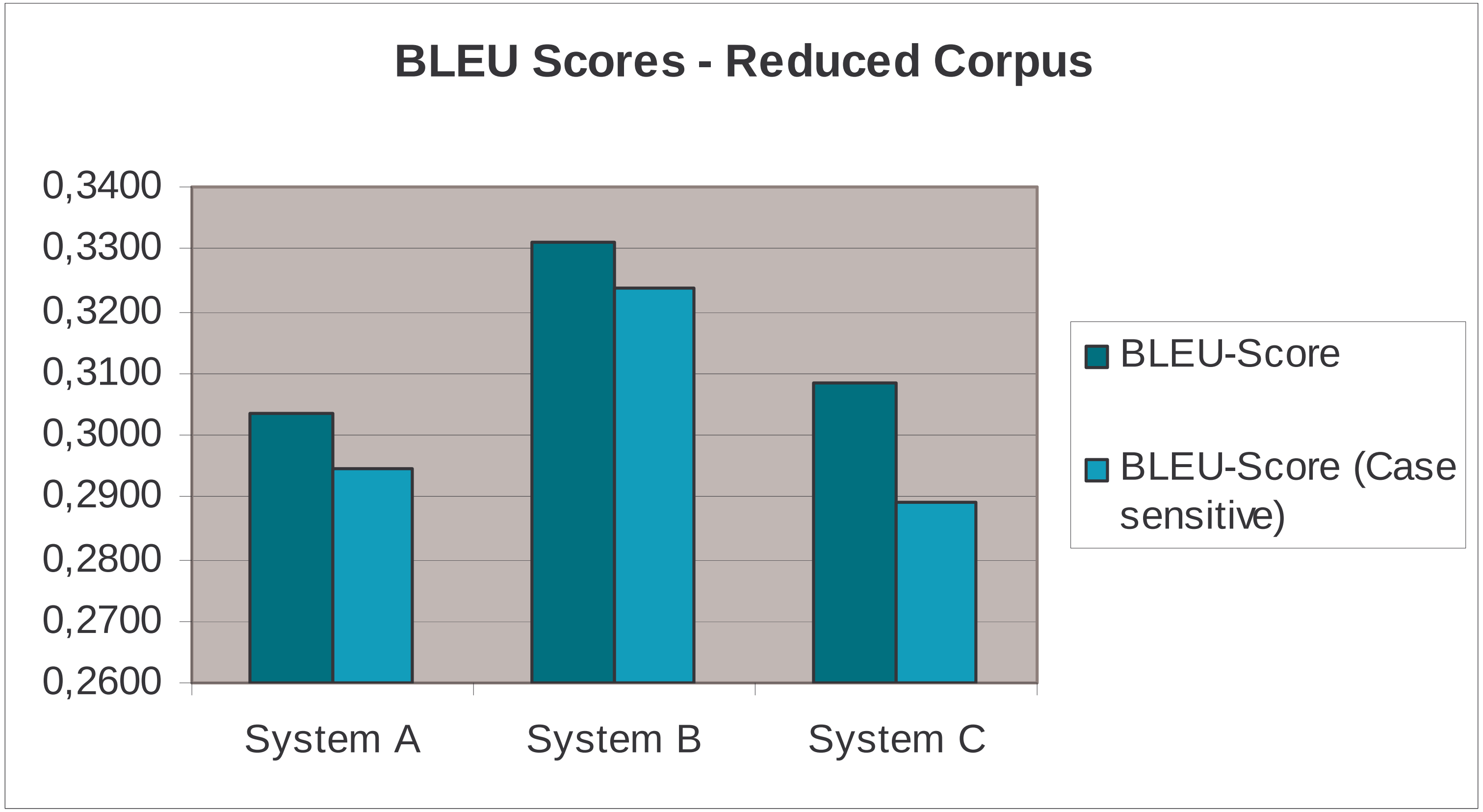
Interpretation of Results
Whole Corpus

- NIST
  - Results of systems B and C are close together, though B leads the classification.
  - The case-sensitive analysis stresses the differences between all systems
  - System A clearly falls behind in both cases
- BLEU
  - System B leads the classification.
  - Results of systems A and C are close together, with a slight advantage for A, both for case-sensitive and non case-sensitive analysis

NIST Results
Reduced Corpus



BLEU Results
Reduced Corpus



Interpretation of Results
BLEU Scores

- NIST
  - System B leads the classification
  - System C follows, closely followed by system A
  - The case sensitive analysis, there is a classification switch between systems A and C (now system C is behind)

- BLEU
  - System B leads the classification
  - System C follows, closely followed by system A
  - The case sensitive analysis, there is a classification switch between systems A and C (now system C is behind)

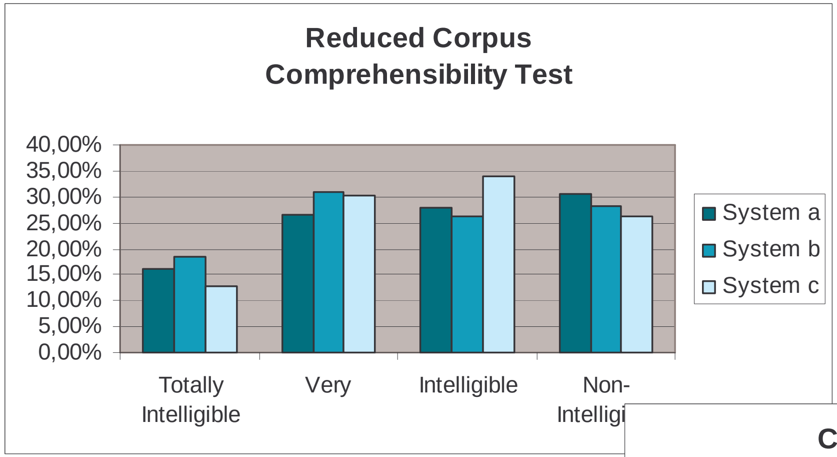
Conclusions

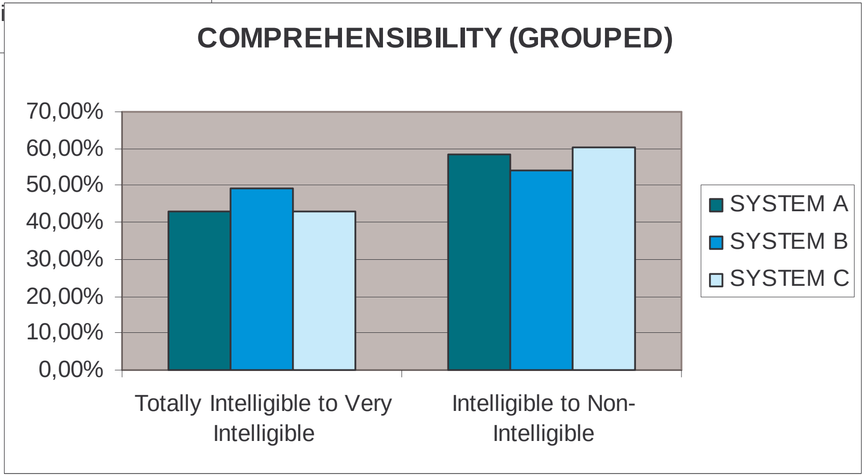
- Clear advantage of system B in all cases and for all scores
- Unclear scores for A and C
- Difficult to state what these results mean for a real translation workflow

Human Evaluation
Reduced Corpus

- Evaluation of following criteria:
  - Comprehensibility: 4-point Scale from “Very Intelligible” to “Non-Intelligible”
  - Post-Editability: 4-point scale from “No post-edition needed” to “Total post-edition”
- Properties of criteria (based on Rodrigo & Braun Chen 01 and derived from FEMTI)
  - K4IN: Key for Information Purposes -> Comprehensibility
  - K4TR: Key for Dissemination Purposes -> Post-Editability

Human Evaluation Comprehensibility Results

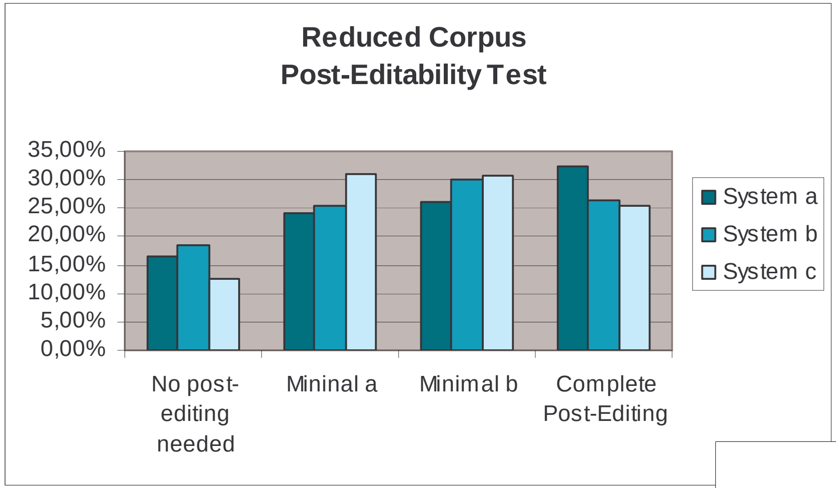


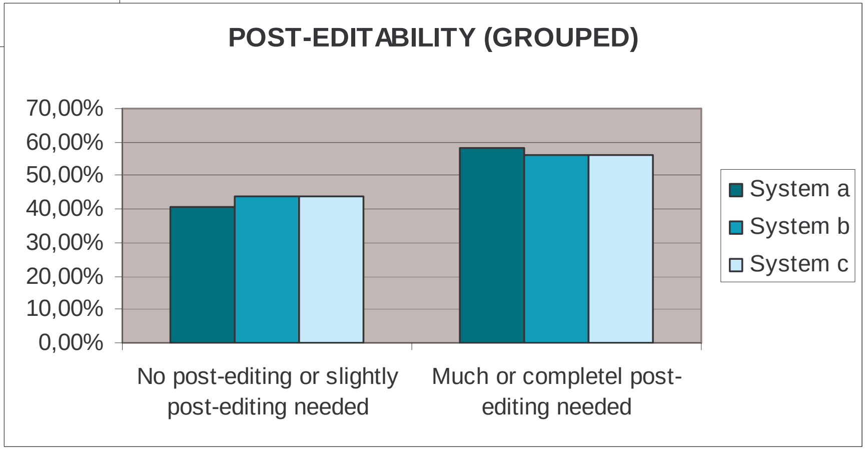


Interpretation of Results Comprehensibility

- System B leads in the categories “Totally and very intelligible” and occupies a middle range in the “non-intelligible” category
- System A has a middle score in “Totally intelligible”, but a high score in “no intelligible”;
- System C has a middle score in “very intelligible”, and the highest scores in “intelligible” as well as the lowest in “non-intelligible”.
  - Assumption: improvement of middle scores by implementing imperative construction rule (German -> English)

Human Evaluation Post-Editability Results





Human Evaluation
Post-Editability Results

- System A offers the highest number of total-postedition and, despite the middle range in “no post-edition”, the low score in minimal post-edition makes it fall behind
- System B offers the highest result in “non post-edition needed” and middle results in the rest categories
- System C offers the lowest no post-edition needed result, but also the lowest “total post-edition”, as well as the highest minimal post-edition results.
    - Assumption: improvement of “total post-edition” scores by implementing imperative construction rule (German -> English)

Conclusions

- System A does not offer the desired output quality and falls behind systems B and C. This can be clearly seen both in the human evaluation and in the automatic evaluation.
- System B offers the best overall results, both in the human evaluation and in the automatic evaluation.
- Systems C offers middle results, though sometimes better results than the other two systems. This is especially significant in the human evaluation of post-editability, where results of B and C are very close together.
  - New Hypothesis: implementation of new grammar rule (imperative structure rule German into English) could improve the quality of system C
    - Trennschloss entriegeln ->**R**elease belt lock
    - Vs
    - Trennschloss entriegeln ->**B**elt lock **r**elease

Outlook

- Optional: Prove hypothesis with system C

- Evaluation of the CL Checker MULTILINT
  - Translation of texts conforming to CL vs. non-conforming texts.
  - Analysis of MULTILINT rules to assess degree of translatability
  - Comparison of rules for human and for machine translatability
  - Study which new rules could improve machine translatability
  - Task-performance evaluation