Our poster: MT-LRE 1 (ID: 636) on Apr 21

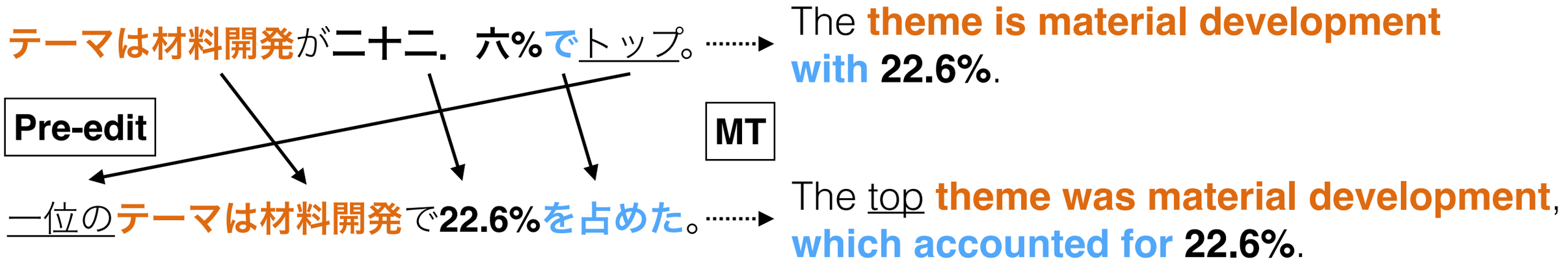
EACL 2021
19–23 April, 2021

Understanding Pre-Editing for Black-Box Neural Machine Translation

Rei Miyata (Nagoya University)
Atsushi Fujita (NICT)

Overview

- **Pre-editing** is a well-tried practice of using MT
  - Controlled language rules for RBMT and SMT [O'Brien 2003; Nyberg+ 2003; Hartley+ 2012]
  - … but not fully studied particularly for NMT [Hiraoka & Yamada 2019; Mehta+ 2020]
- This presentation explains
  1. our methodology for collecting fine-grained pre-editing instances
  2. our extensive analyses of the collected instances
- We revealed
  - potential of pre-editing and NMT approach
  - possible and effective pre-editing operation types
  - general tendencies of impact of pre-editing on NMT output



1. Protocol

Incrementally and minimally rewrite an ST on a trial-and-error basis with the aim of obtaining better MT output [Miyata & Fujita 2017]

No	Source Sentence (Original)	Machine Translation	MEMO for the Unit
8	一位のテーマは材料開発で22．6%を占めた。	The top theme was material development, which accounted for 22.6%.	save

Complete = **satisfactory** output is obtained
(information is fully conveyed without grammatical errors)

Give up = satisfactory output is unlikely to be obtained

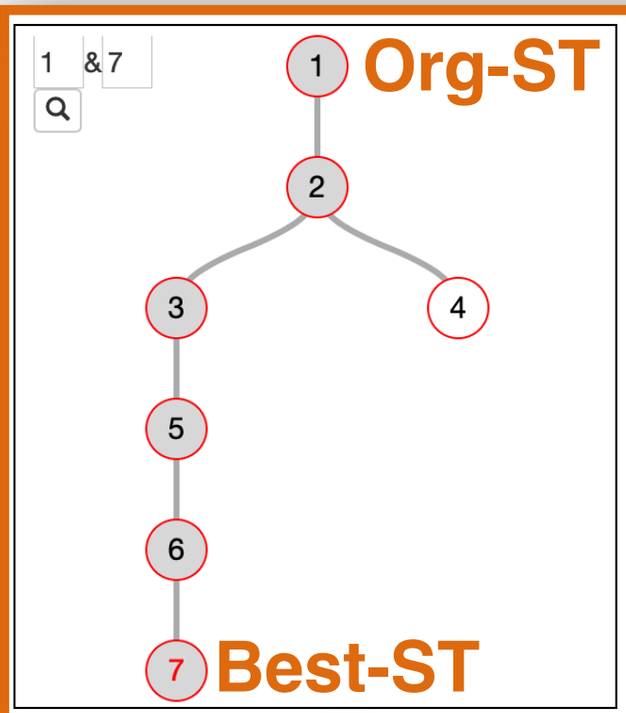
Translate | Complete | Back to Unit List

Pre-editing History

No	Past Sentence (Tree View)	Machine Translation	Memo
★ 7	一位のテーマは材料開発で22．6%を占めた。	The top theme was material development, which accounted for 22.6%.	
☆ 6	一位のテーマは材料開発で二十二．六%を占めた。	The theme of the first place occupied 22. 6% in the material development.	
☆ 5	一位のテーマは材料開発で二十二．六%。	The number one theme is the development of materials, which is 22. 6%.	
☆ 4	テーマの一位は材料開発で二十二．六%。	The top theme was the development of materials, with 22. 6%.	
☆ 3	トップのテーマは材料開発で二十二．六%。	The top theme is 22. 6% in material development.	
☆ 2	テーマのトップは材料開発で二十二．六%。	The top theme is material development at 22. 6%.	
☆ 1	テーマは材料開発が二十二．六%とトップ。	The theme is material development with 22. 6%.	

1. Protocol

Incrementally and minimally rewrite an ST on a trial-and-error basis with the aim of obtaining better MT output [Miyata & Fujita 2017]



	テーマは材料開発が二十二．六%とトップ。	The theme is material development with 22. 6%.
	テーマのトップは材料開発で二十二．六%。	The top theme is material development at 22. 6%.
	トップのテーマは材料開発で二十二．六%。	The top theme is 22. 6% in material development.
	一位のテーマは材料開発で二十二．六%。	The number one theme is the development of materials, which is 22. 6%.
	一位のテーマは材料開発で二十二．六%を占めた。	The theme of the first place occupied 22. 6% in the material development.
	一位のテーマは材料開発で22．6%を占めた。	The top theme was material development, which accounted for 22.6%.

No

8

Pre-editing History

Tree View

No	Past Sentence		Machine Translation	Memo
★ 7	一位のテーマは材料開発で22．6%を占めた。	➜	The top theme was material development, which accounted for 22.6%.	
☆ 6	一位のテーマは材料開発で二十二．六%を占めた。	➜	The theme of the first place occupied 22. 6% in the material development.	
☆ 5	一位のテーマは材料開発で二十二．六%。	➜	The number one theme is the development of materials, which is 22. 6%.	
☆ 4	テーマの一位は材料開発で二十二．六%。	➜	The top theme was the development of materials, with 22. 6%.	
☆ 3	トップのテーマは材料開発で二十二．六%。	➜	The top theme is 22. 6% in material development.	
☆ 2	テーマのトップは材料開発で二十二．六%。	➜	The top theme is material development at 22. 6%.	
☆ 1	テーマは材料開発が二十二．六%とトップ。	➜	The theme is material development with 22. 6%.	

Implementation

- Condition
  - **3 translation directions:** Ja-En, Ja-Zh, Ja-Ko
  - **2 MT systems:** Google, TexTra
  - **4 text domains:** hospital conversation, municipal procedure, Japanese-origin news article, English-origin news article
- Data
  - 100 Japanese source sentences (**Org-ST,** 25 sentences × 4 domains)
  - for all of the 6 settings (3 translation directions × 2 MT systems)
- Editor
  - One **professional translator** for each translation direction
    - each worked for 100 Org-STs × 2 MT systems

Collected pre-editing instances

		# instances		# units	
Lang.	**System**	**Total**	**Avg.**	**Org=Satisfactory**	**Best=Satisfactory**
Ja-En	**Google**	1332	13.32	27/100	98/100
	Textra	1260	12.60	23/100	96/100
Ja-Zh	**Google**	1371	13.71	2/100	91/100
	Textra	812	8.12	8/100	94/100
Ja-Ko	**Google**	950	9.50	1/100	96/100
	Textra	927	9.27	8/100	96/100
Total		6652	11.09	69/600	571/600

✓ Potential of pre-editing + NMT

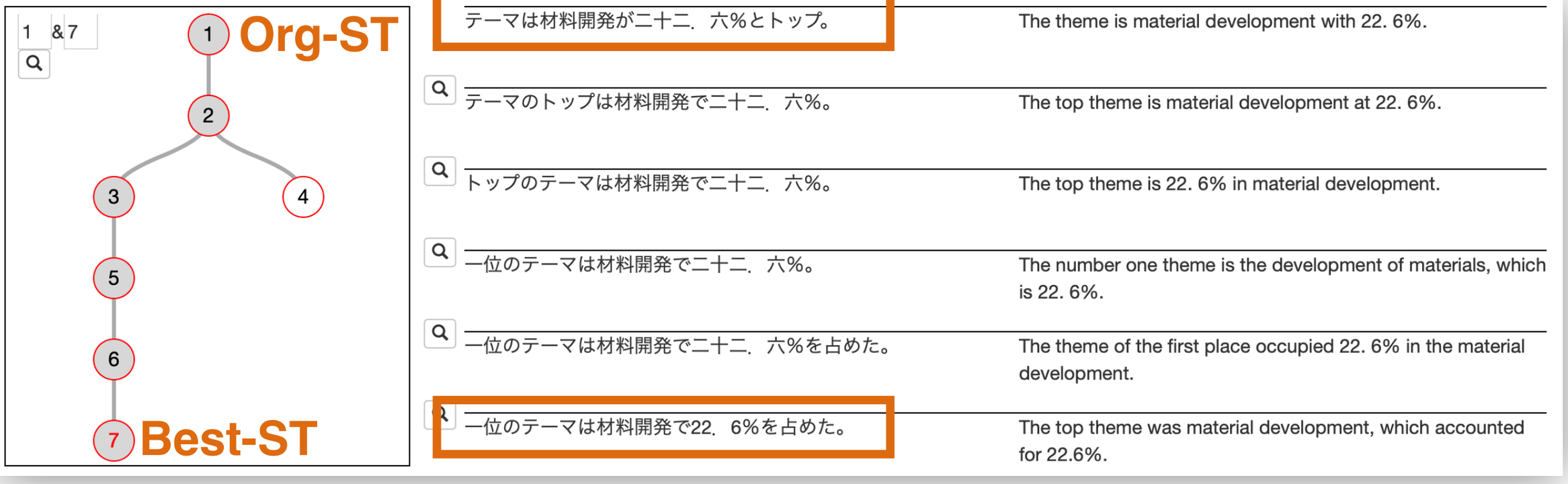
2. Analyses of the collected instances

a. Characteristics of pre-edited sentences

b. Diversity of pre-editing operations

c. Impact of pre-editing on NMT

a. Characteristics of pre-edited sentences



- Data: 100 Org-ST vs. 100 Best-ST (for each of the six conditions)
- Viewpoints
  - Structural characteristics: length, attachment distance, dependency depth
  - Lexical characteristics: Type/Token Ratio (TTR), frequency rank in Wikipedia

a. Characteristics of pre-edited sentences

	Org	Best					
		Ja-En		Ja-Zh		Ja-Ko	
		Google	Textra	Google	Textra	Google	Textra
Sentence length	25.4	27.8	26.9	28.6	27.1	27.8	26.9
Attachment distance	1.95	1.97	1.99	1.99	1.99	2.00	1.98
Dependency depth	3.57	3.73	3.68	3.73	3.77	3.78	3.76
Type/Token Ratio	0.398	0.386	0.395	0.387	0.392	0.384	0.392
Word freq. rank (Med.)	170	143	154	143	155	143	169.5

✓ The structural complexity generally increased
✓ Low-frequency words tended to be avoided

a. Characteristics of pre-edited sentences

	Org	Best					
		Ja-En		Ja-Zh		Ja-Ko	
		Google	Textra	Google	Textra	Google	Textra
Sentence length	25.4	27.8	26.9	28.6	27.1	27.8	26.9
Attachment distance	1.95	1.97	1.99	1.99	1.99	2.00	1.98
Dependency depth	3.57	3.73	3.68	3.73	3.77	3.78	3.76
Type/Token Ratio	0.398	0.386	0.395	0.387	0.392	0.384	0.392
Word freq. rank (Med.)	170	143	154	143	155	143	169.5

✓ **The structural complexity generally increased**

✓ **Low-frequency words tended to be avoided**

a. Characteristics of pre-edited sentences

	Org	Best					
		Ja-En		Ja-Zh		Ja-Ko	
		Google	Textra	Google	Textra	Google	Textra
Sentence length	25.4	27.8	26.9	28.6	27.1	27.8	26.9
Attachment distance	1.95	1.97	1.99	1.99	1.99	2.00	1.98
Dependency depth	3.57	3.73	3.68	3.73	3.77	3.78	3.76
Type/Token Ratio	0.398	0.386	0.395	0.387	0.392	0.384	0.392
Word freq. rank (Med.)	170	143	154	143	155	143	169.5

✓ **The structural complexity generally increased**

✓ **Low-frequency words tended to be avoided**

b. Pre-editing operation types [cf. Miyata & Fujita 2017]

ID	#	Structure
S01	14	Sentence splitting
S02	27	Structural change
S03	19	Use/disuse of topicalisation
S04	14	Insertion of subject/object
S05	12	Use/disuse of clause-ending noun
S06	4	Change of voice
S07	5	Other structural changes
ID	**#**	**Phrase**
P01	71	Insertion/deletion of punctuation
P02	28	Use/disuse of chunking marker(s)
P03	31	Phrase reordering
P04	7	Change of modification
P05	40	Change of connective expression
P06	36	Change of parallel expression
P07	16	Change of apposition expression
P08	13	Change of noun/verb phrase
P09	28	Use/disuse of compound noun
P10	17	Use/disuse of affix
P11	5	Change of sahen noun expression
P12	9	Change of formal noun expression
P13	5	Other phrasal changes

ID	#	Content word
C01	118	Use of synonymous words
C02	21	Use/disuse of abbreviation
C03	14	Use/disuse of anaphoric expression
C04	11	Use/disuse of emphatic expression
C05	30	Category indication/suppression
C06	10	Explanatory paraphrase
C07	94	Change of content
ID	**#**	**Functional word**
F01	47	Change of particle
F02	31	Change of compound particle
F03	12	Change of aspect
F04	4	Change of tense
F05	11	Change of modality
F06	10	Use/disuse of honorific expression
ID	**#**	**Orthography**
O01	61	Japanese orthographical change
O02	16	Change of half-/full-width character
O03	2	Insertion/deletion/change of symbol
O04	4	Other orthographical change
ID	**#**	**Error**
E01	22	Grammatical errors
E02	16	Content errors

✓ **Sentence splitting was not frequently used**

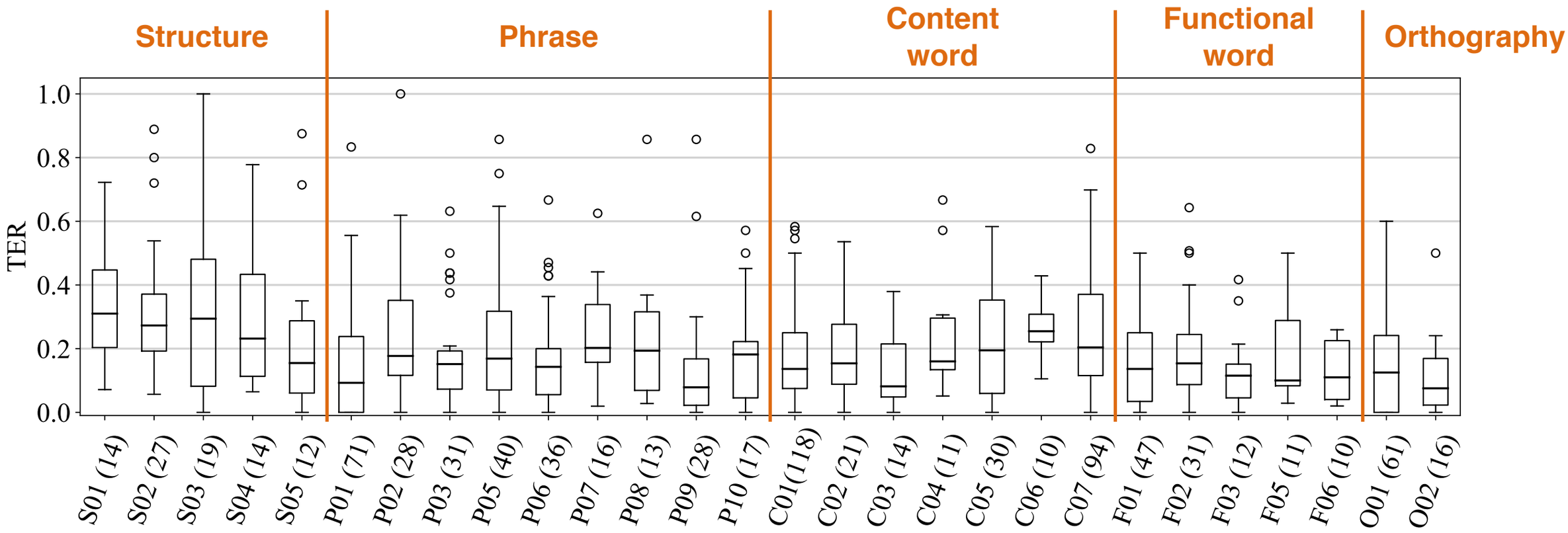
b. Strategies for effective pre-editing

- Informational strategies [e.g., Vinay & Darbelnet 1958; Chesterman 1997]
  - **Explicitation**, Implicitation, Preservation

Explicitation Strategy	#	Example of ST pre-editing	MT output
Information addition	142	12日は台湾の休日のため休場。	The twelfth is a holiday in Taiwan.
		→ 12日は台湾の休日のため**株式市場は**休場。	**The stock market** was closed on the twelfth due to a holiday in Taiwan.
Use of clear relation	103	来院しなくても10日前後で登録のクレジットカードから**引き落としを行います**。	**Withdraw from your registered credit card** in about 10 days without visiting the hospital.
		→ 来院しなくても10日前後で登録のクレジットカードから**引き落としが行われます。**	Even if you do not visit the hospital, **your credit card will be debited** in about 10 days.
Use of narrower sense	54	採尿と採便を**出して**ください。	Please **collect** urine and feces.
		→ 採尿と採便を**提出して**ください。	Please **submit** urine and stool samples.
Normalisation	30	単位は億円。	Figures are in **billions of yen**.
		→ 単位は億円**です**。	The unit is **100 million yen**.

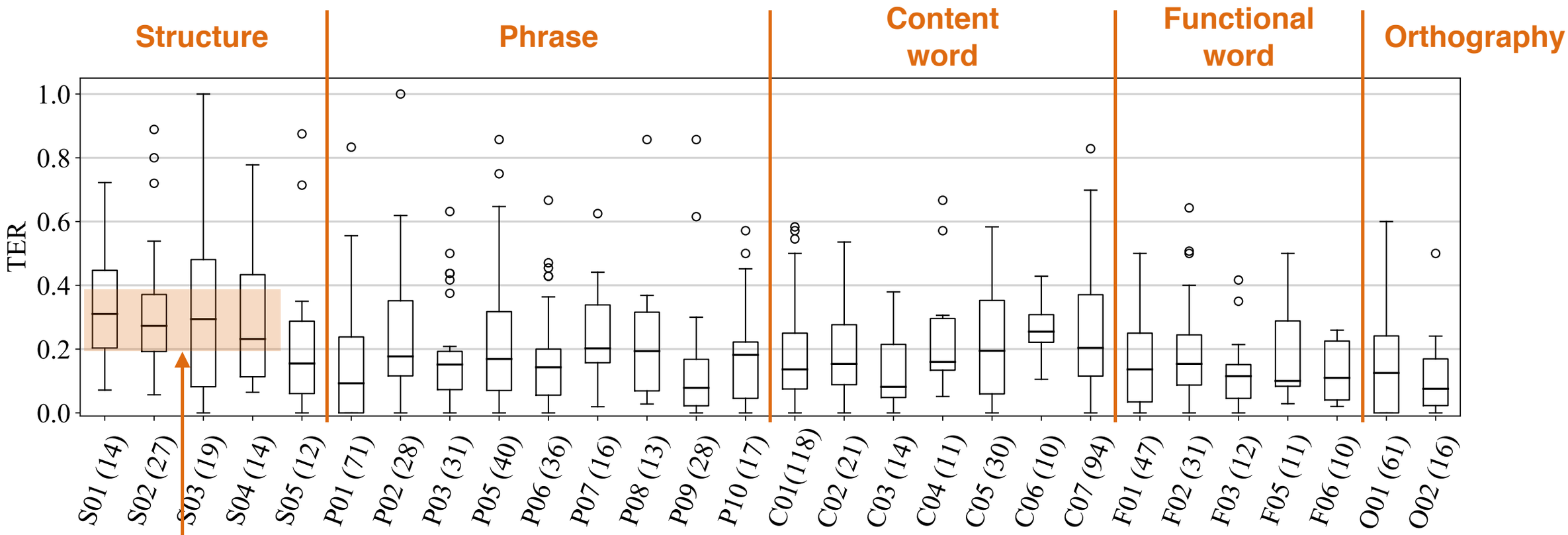
✓ Explicitation is the key to the effective pre-editing

c. Impact of editing operations on MT



✓ **Structural edits in ST tend to cause major changes in MT**
✓ **Rewriting functional words and orthography seldom impacts**
✓ **Phrase reordering in ST does not affect MT output much**

c. Impact of editing operations on MT

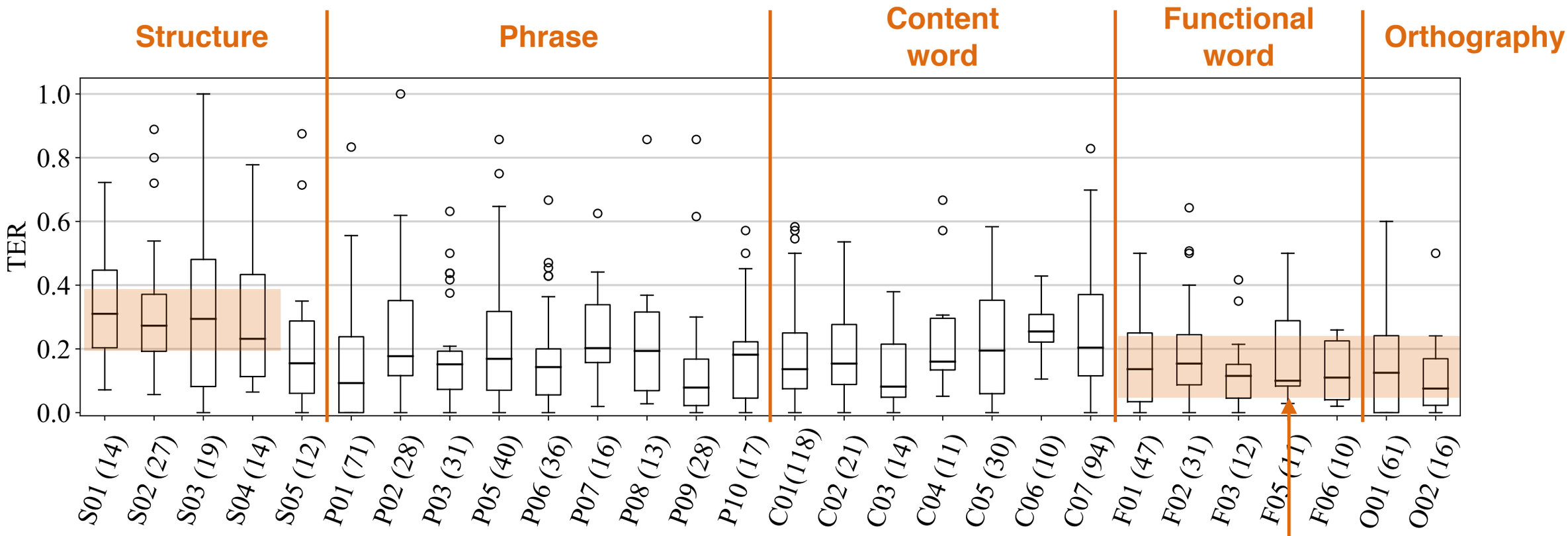


✓ **Structural edits in ST tend to cause major changes in MT**

✓ **Rewriting functional words and orthography seldom impacts**

✓ **Phrase reordering in ST does not affect MT output much**

c. Impact of editing operations on MT

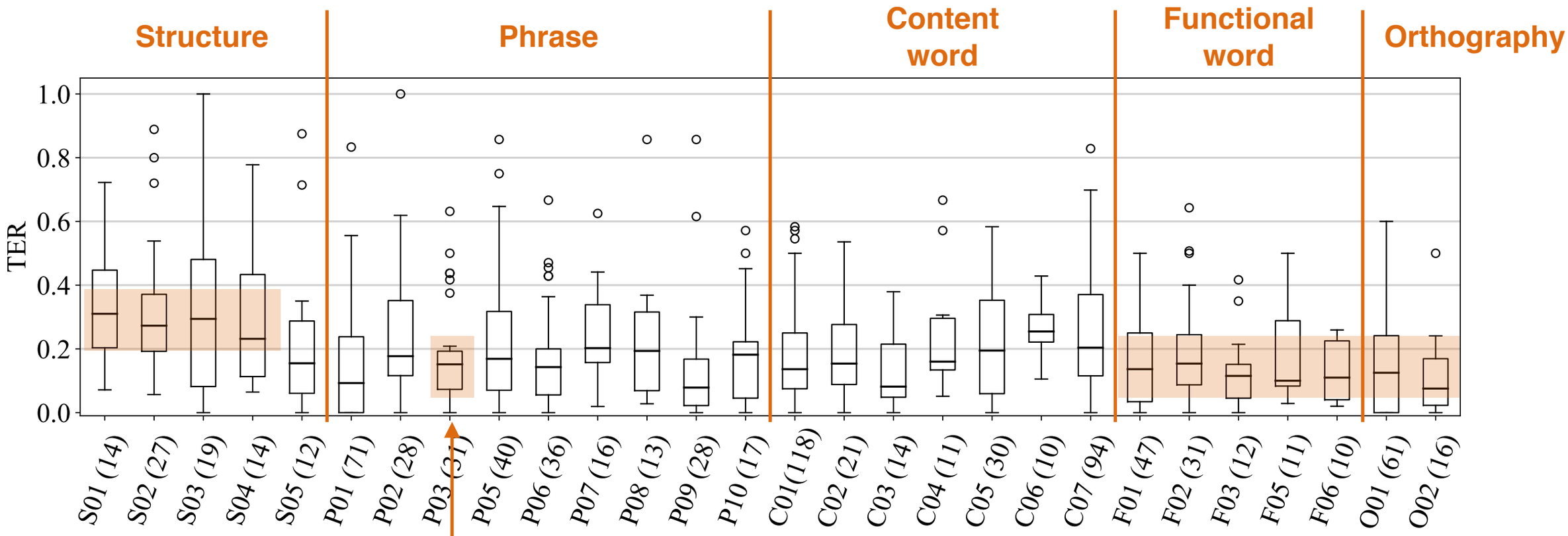


✓ **Structural edits in ST tend to cause major changes in MT**

✓ **Rewriting functional words and orthography seldom impacts**

✓ **Phrase reordering in ST does not affect MT output much**

c. Impact of editing operations on MT



✓ **Structural edits in ST tend to cause major changes in MT**

✓ **Rewriting functional words and orthography seldom impacts**

✓ **Phrase reordering in ST does not affect MT output much**

Summary

1. 6,652 manual pre-editing instances were collected
   - across multiple translation directions, MT systems, and domains
   - **demonstrating the potential of pre-editing + NMT**
2. Detailed analyses revealed:
   - **Enhancing the explicitness of an ST is more important** than making the ST shorter and simpler
   - Although NMT is generally unpredictable, **there are recognisable tendencies of impact of pre-editing on NMT**

Future research directions

1. Implementation of an automatic pre-editor for black-box NMT scenarios
2. Development of tools to support human pre-editors

Our poster: MT-LRE 1 (ID: 636) on Apr 21