# Enhancing Normalization of Ukrainian News Texts

## Information Systems and Technologies (IST-2025), May 19-22, 2025, Kharkiv, Ukraine

Danylo O. **Horielov**
Oleksander V. **Vechur**

Kharkiv National University of Radio Electronics, Ukraine

# State of the problem

## Problem:

Ukrainian news texts contain inconsistent formatting patterns for: phone numbers, quotation marks, and apostrophes

## Related researches:

- Aliero - Analysis of latest normalization approaches
- Starko - Text preprocessing system for GRAC corpus
- Vakulenko - Normalization focused on text-to-speech conversion
- Chaplynskyi - UberText 2.0 with basic news normalization and cleaning
- Cudak - Address phone number normalization as part of their research on POI matching
- Goot and Çetinoğlu  - Address the challenge of lexical normalization in code-switched data

# Problem 1: Phone Number Inconsistency

## There are a lot of variations to write a phone number:

+38 0XX XXX XX XX; +380 XXX XXX-XXX, … , +38 (0XX) XX-XX-XXX

## These inconsistencies are prevalent across news publications and include:

- Inconsistent representation of country codes and parentheses
- Varied grouping of digits
- Different delimiters (spaces, hyphens)
- Different formats for special numbers (like toll-free numbers)

# Phone Number Normalization Approach

## Standard format requirements

**The recommended format for phone numbers is:**

+380 (XX) XXX-XX-XX

We propose an additional format for special numbers:

+380 (XXX) XX-XX-XX

Because "+380 (800) XX..." is more recognizable then "+380 (80) 0X…"

# Phone Number Normalization Approach

## Phone pattern.

1. We limit normalization to Ukrainian phone numbers

2. We identified several groups of typical phone number patterns:

   AA_XXX_XX_XX, AA_XX_XX_XXX, AA_XX_XXX_XX, AAX_XX_XX_XX,
   AAXX_X_XX_XX, AAA_XXX_XXX, AAA_XX_XX_XX

3. Within these groups we looking for this prefixes:

   +380 AA; +38 0AA; +38 (0AA); +380 (AA); 0AA; (0AA)

4. Avoid substringing with other numbers

There are total of 24 regular expression

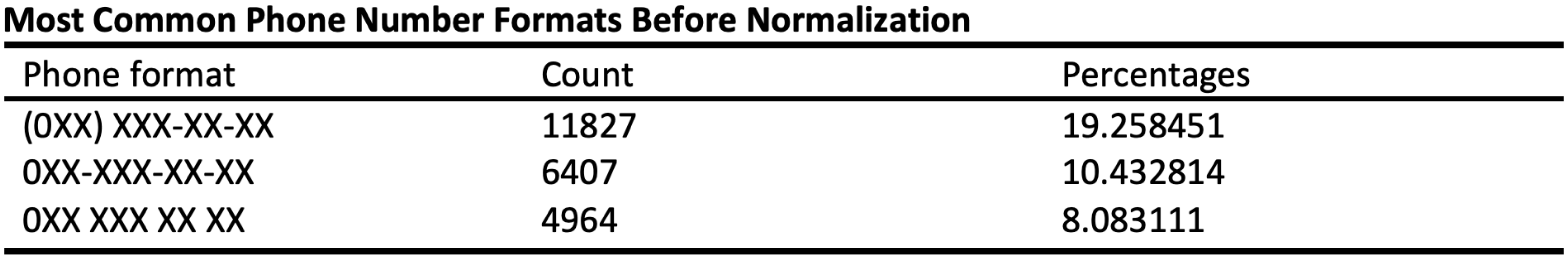
# Experiment Setup

To evaluate our normalization methods, we conducted experiments using the Uber Text 2.0 News Cleaned subcorpus. The corpus structure consists of news articles separated by multiple empty lines

We processed data by iterating through the corpus, accumulating lines until reaching delimiter lines, and then treating each accumulated block as a single news article for processing.

# Phone Number Normalization Result

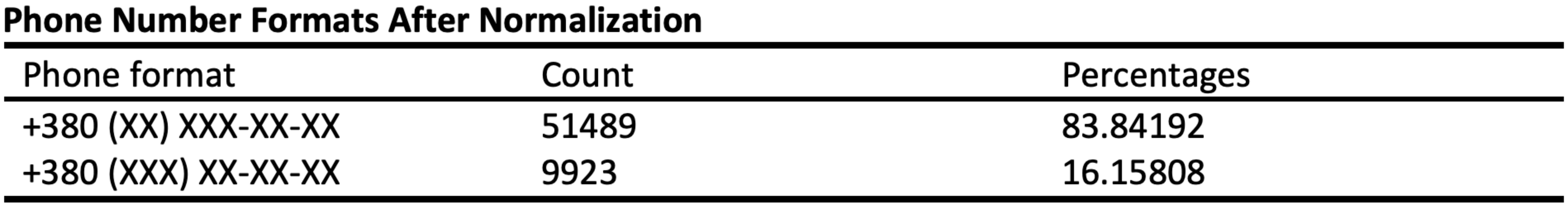
Our analysis identified 302 unique phone number formats in the UT2 corpus



**Most Common Phone Number Formats Before Normalization**

| Phone format | Count | Percentages |
|---|---|---|
| (0XX) XXX-XX-XX | 11827 | 19.258451 |
| 0XX-XXX-XX-XX | 6407 | 10.432814 |
| 0XX XXX XX XX | 4964 | 8.083111 |

We successfully consolidated all 302 formats into standardized versions



**Phone Number Formats After Normalization**

| Phone format | Count | Percentages |
|---|---|---|
| +380 (XX) XXX-XX-XX | 51489 | 83.84192 |
| +380 (XXX) XX-XX-XX | 9923 | 16.15808 |

# Problem 2: Apostrophe Inconsistency

## Various similar-looking symbols cause inconsistencies:

'  ´ ‘ ’ ʻ ʼ `

## Example from real news:

Новий прем’єр-міністр Канади Марк Карні складе присяги як 24-й очільник уряду країни в п'ятницю, 14 березня

Apostrophes are sometimes misused as quotation marks.

# Apostrophe Normalization Approach

## We recommend following standard:

U+02BC - ’

## Why this choice?

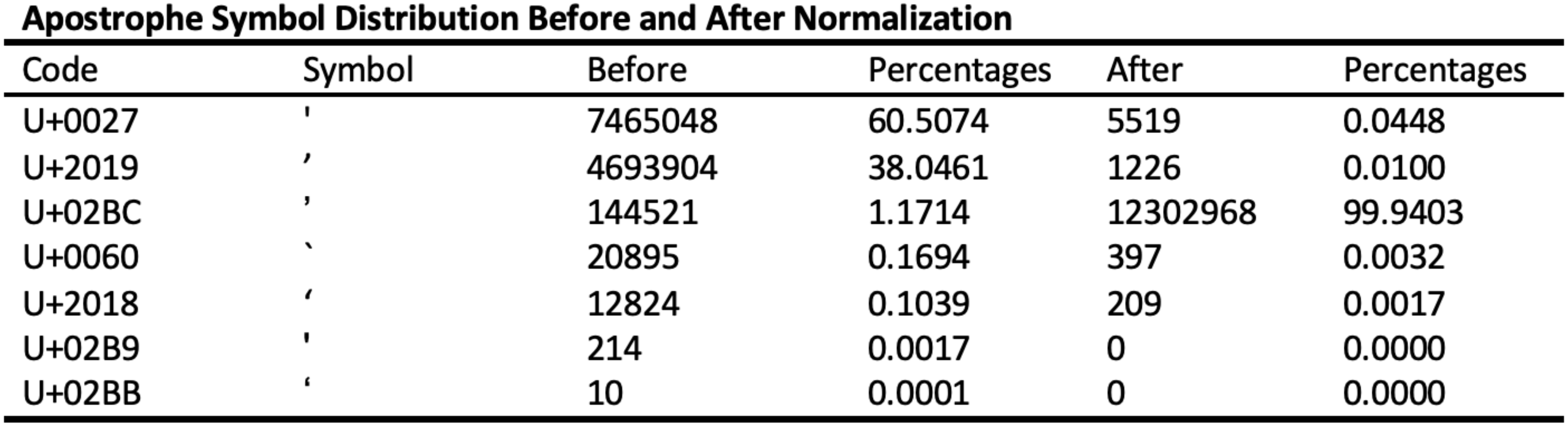
- Default on Apple devices when using the Ukrainian keyboard
- Curved shape matches the traditional handwritten apostrophe in Ukrainian
- Visually distinct from quotation marks, reducing ambiguity

# Apostrophe Normalization Approach

## Normalization steps:

1. Double Occurrences – Remove or replace duplicated apostrophes
2. Word-Internal Apostrophes – Ensure correct usage within words
3. Boundary Cases – Handle apostrophes at the beginning or end of words
4. Ambiguous Cases – Resolve confusion with similar-looking symbols
5. Warning and Error Handling – Detect and flag problematic cases

# Apostrophe Normalization Result



**Apostrophe Symbol Distribution Before and After Normalization**

| Code | Symbol | Before | Percentages | After | Percentages |
|---|---|---|---|---|---|
| U+0027 | ' | 7465048 | 60.5074 | 5519 | 0.0448 |
| U+2019 | ’ | 4693904 | 38.0461 | 1226 | 0.0100 |
| U+02BC | ʼ | 144521 | 1.1714 | 12302968 | 99.9403 |
| U+0060 | ` | 20895 | 0.1694 | 397 | 0.0032 |
| U+2018 | ‘ | 12824 | 0.1039 | 209 | 0.0017 |
| U+02B9 | ʹ | 214 | 0.0017 | 0 | 0.0000 |
| U+02BB | ʻ | 10 | 0.0001 | 0 | 0.0000 |

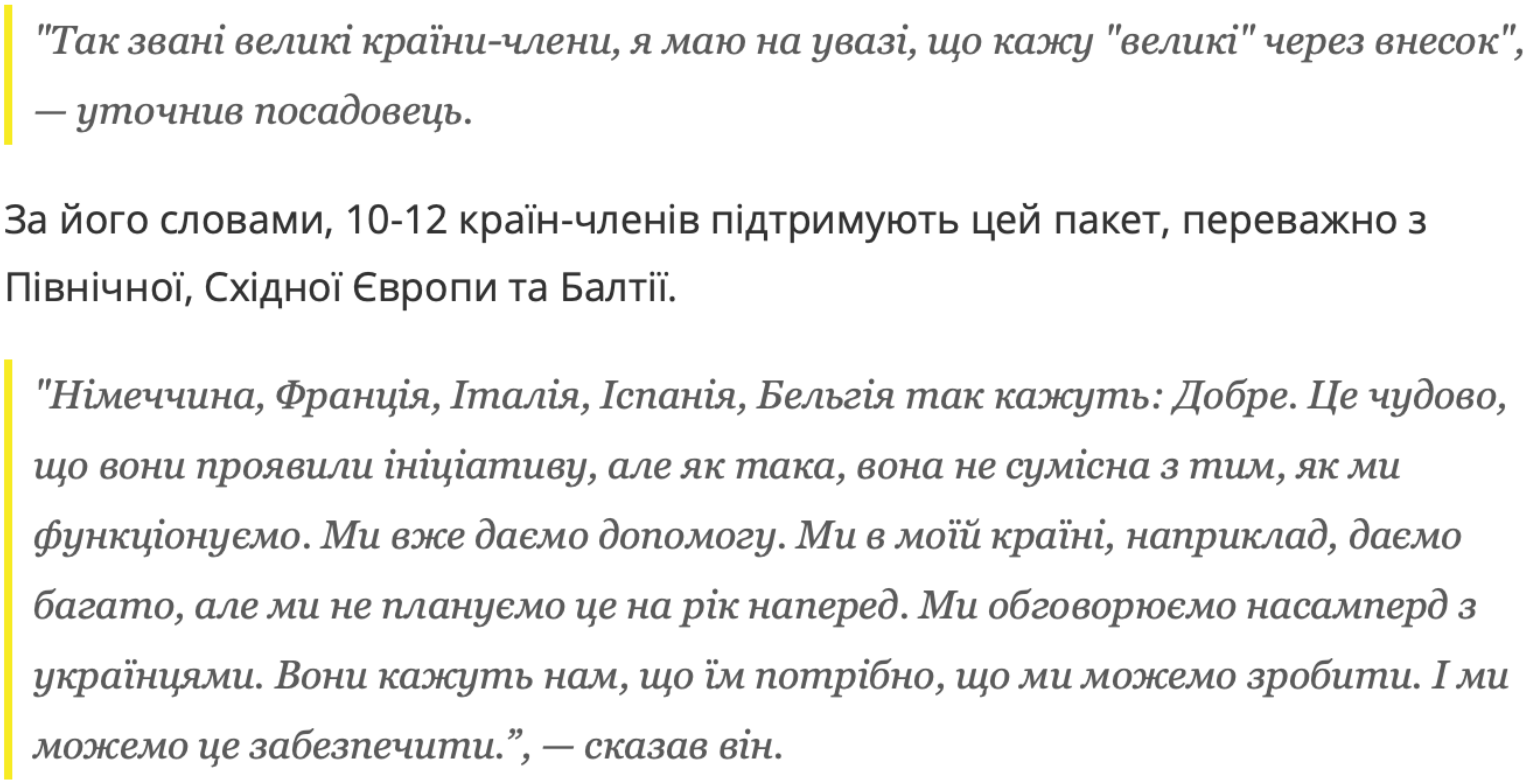
- 99.94% of apostrophes were successfully converted to U+02BC
- 0.06% of symbols remaining in their original form.
- 0.22% of the original apostrophe symbols were identified as quotation

# Problem 3: Quotes Inconsistency

## There are various symbols:

" « » “ ” ‟ „ ❝ ❞

## Example from real news:



> *"Так звані великі країни-члени, я маю на увазі, що кажу "великі" через внесок", — уточнив посадовець.*

За його словами, 10-12 країн-членів підтримують цей пакет, переважно з Північної, Східної Європи та Балтії.

> *"Німеччина, Франція, Італія, Іспанія, Бельгія так кажуть: Добре. Це чудово, що вони проявили ініціативу, але як така, вона не сумісна з тим, як ми функціонуємо. Ми вже даємо допомогу. Ми в моїй країні, наприклад, даємо багато, але ми не плануємо це на рік наперед. Ми обговорюємо насамперд з українцями. Вони кажуть нам, що їм потрібно, що ми можемо зробити. І ми можемо це забезпечити.”, — сказав він.*

# Quotes Normalization Approach

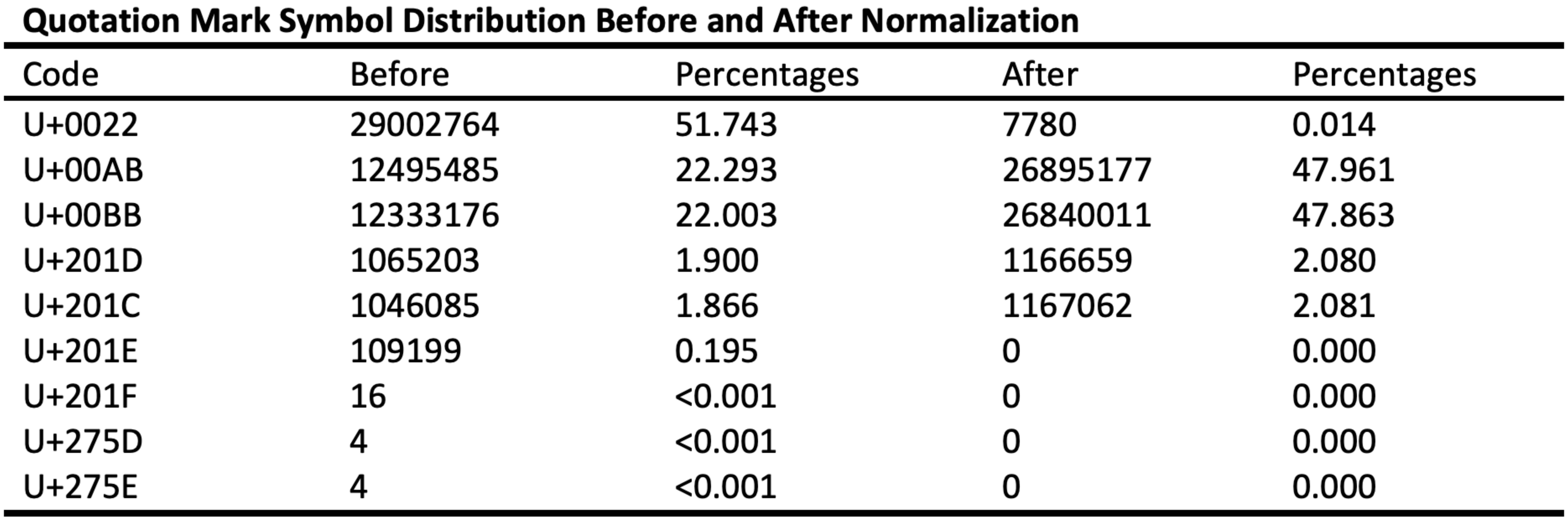
## Standard format requirements:

«Text text “text” text»

## Algorithm consists of three sequential stages:

1. Symbol unification
2. Contextual replacement
3. Nested quotation handling

# Quotes Normalization Result



**Quotation Mark Symbol Distribution Before and After Normalization**

| Code | Before | Percentages | After | Percentages |
|---|---|---|---|---|
| U+0022 | 29002764 | 51.743 | 7780 | 0.014 |
| U+00AB | 12495485 | 22.293 | 26895177 | 47.961 |
| U+00BB | 12333176 | 22.003 | 26840011 | 47.863 |
| U+201D | 1065203 | 1.900 | 1166659 | 2.080 |
| U+201C | 1046085 | 1.866 | 1167062 | 2.081 |
| U+201E | 109199 | 0.195 | 0 | 0.000 |
| U+201F | 16 | <0.001 | 0 | 0.000 |
| U+275D | 4 | <0.001 | 0 | 0.000 |
| U+275E | 4 | <0.001 | 0 | 0.000 |

The balanced percentages between opening and closing quotation marks indicate successful pairing.

# Discussion and Future work

- Phone Number Normalization Limitations
- Text Preprocessing Considerations
- Apostrophe Normalization Challenges

# Conclusion

- For phone numbers, we consolidated 302 different formats into two standardized formats
- Our apostrophe normalization achieved 99.94% conversion to the target standard
- The quotation mark normalization successfully implemented the Ukrainian orthographic standard of using guillemets for primary quotations and curly quotes for nested quotations

# Thank you