# COMPUTATIONAL SUPPORT FOR FINDING WORD CLASSES: A CASE STUDY OF ABUI

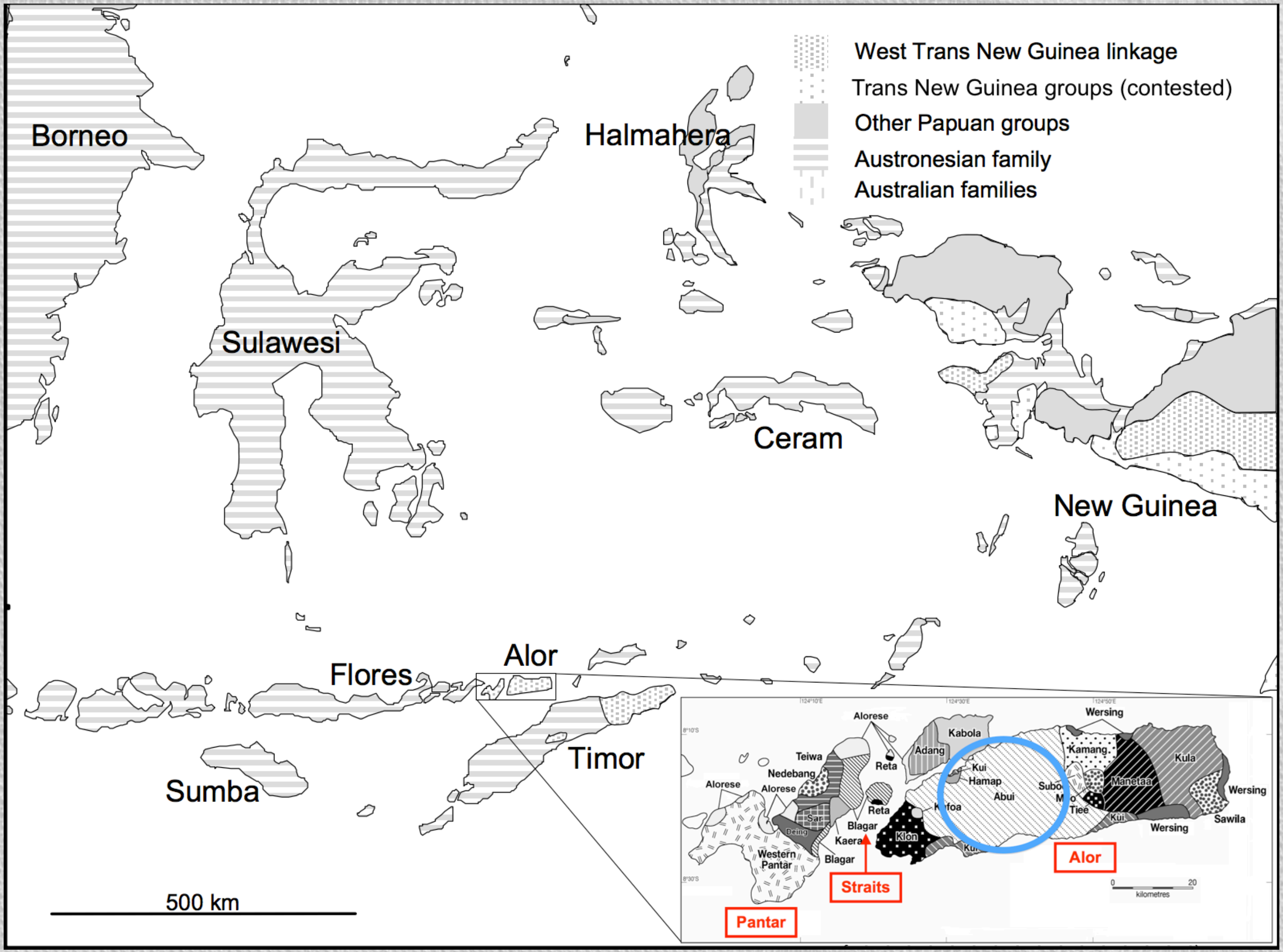
Olga Zamaraeva*, František Kratochvíl**, Emily M. Bender*,

Fei Xia* and Kristen Howell*

*University of Washington, USA

**Nanyang Technological University, Singapore

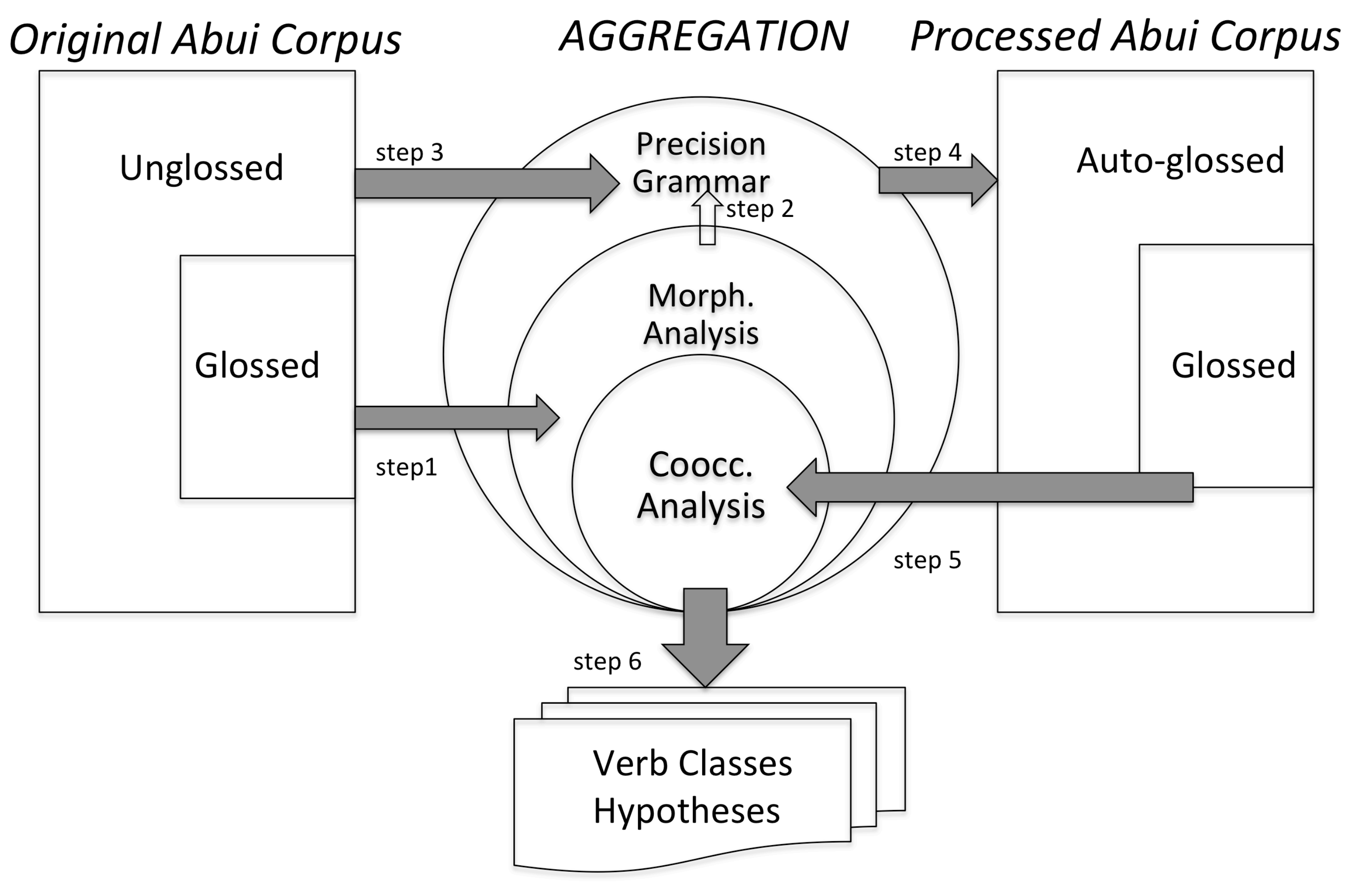
# Computational Overview for a Corpus of Abui [abz]



# Abui Verb Inflectional Classes

- Verbs combine with different prefix series (Conditions I-IV)
- RQ: Are the classes semantically coherent?
- Only ~10% of the corpus is glossed
- We segment and gloss words automatically (as verbs)
  - By applying lexical rules learned from the glossed portion
- Then we look at prefix-verb cooccurrences in the entire corpus
- ...and compare what we find in the collected corpus with a curated (elicited) set

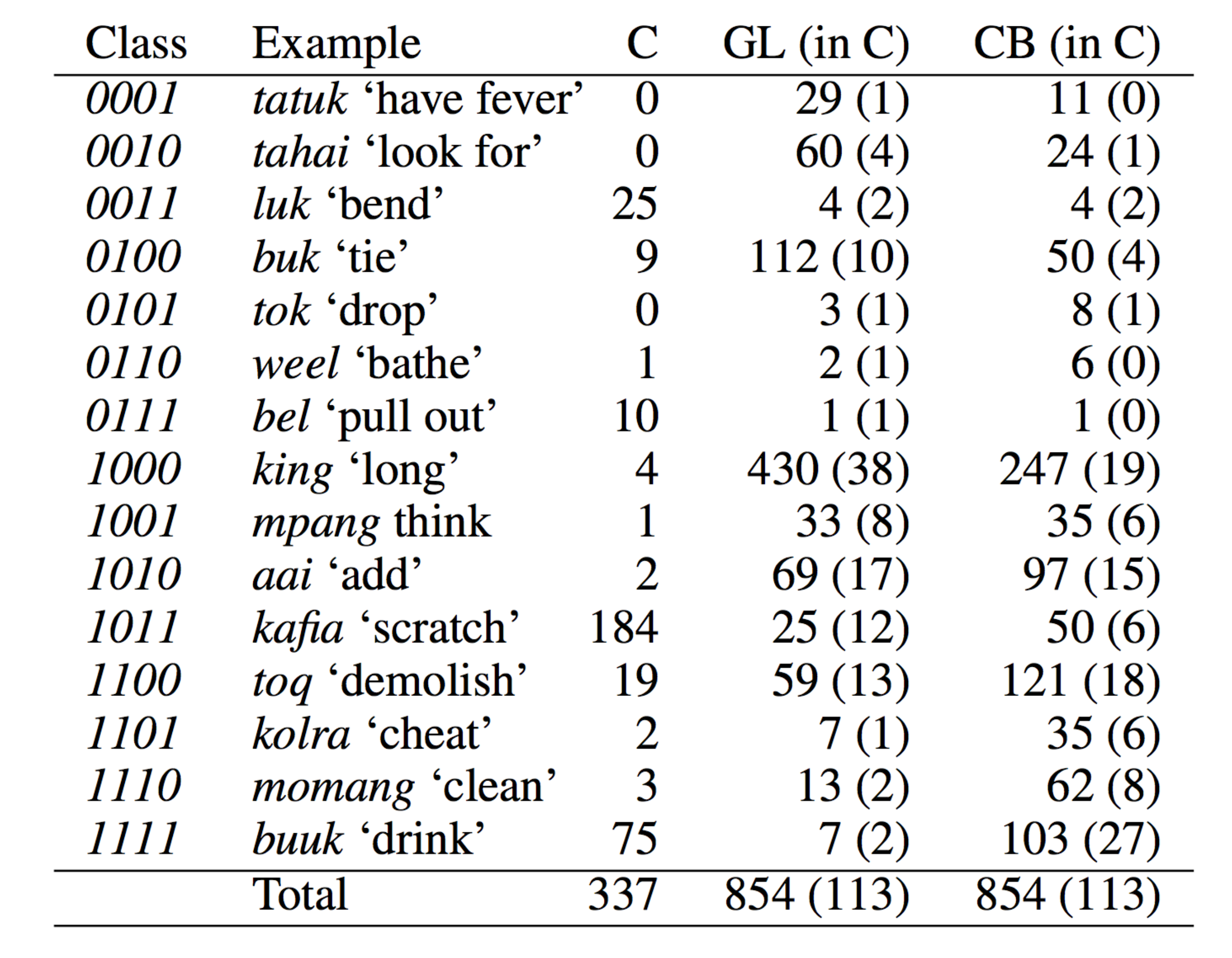
# System overview



# Verbs in the IGT corpus

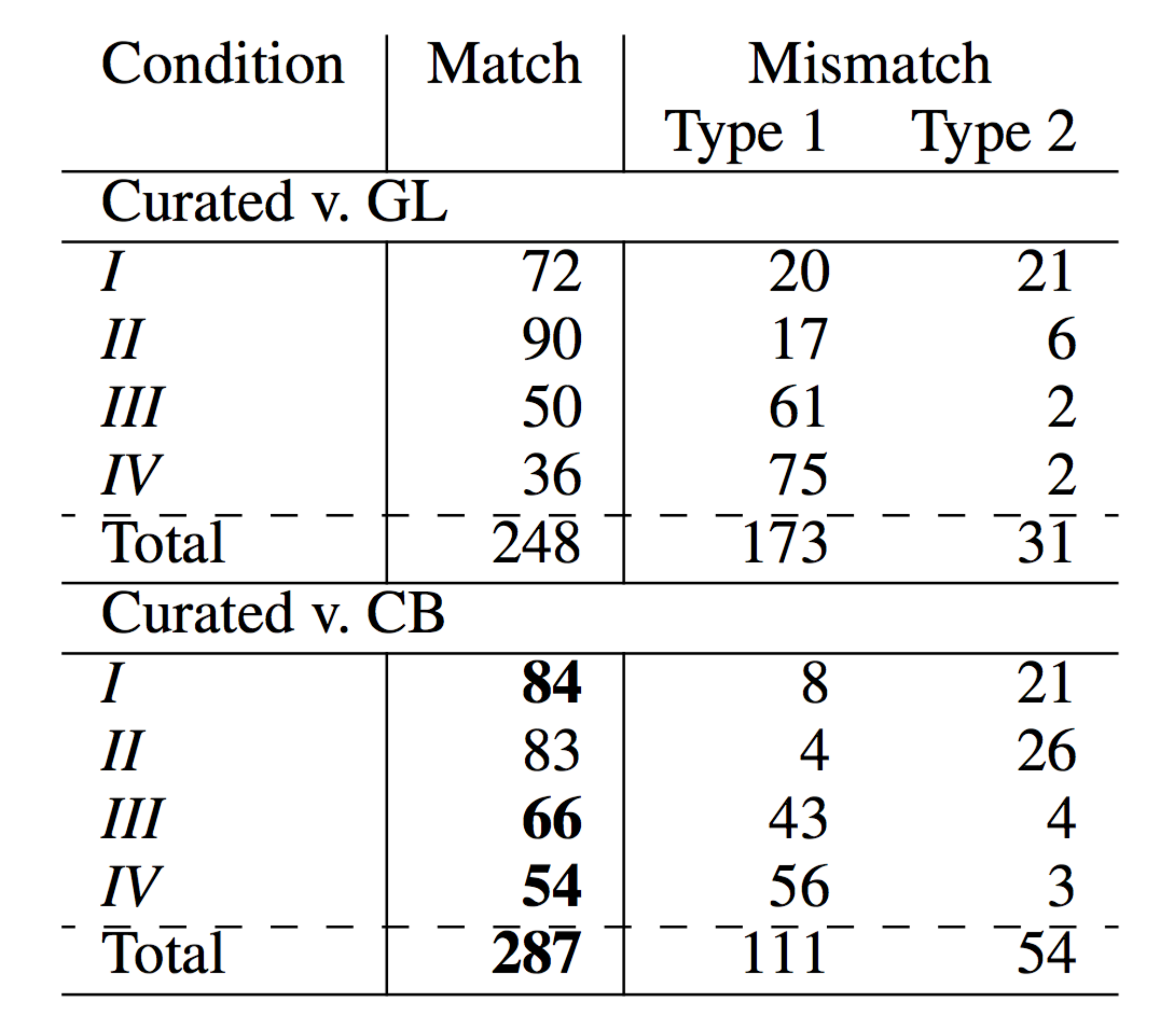
| | Verb tokens | Verb types |
|---|---|---|
| Originally Glossed | 8,609 | 2,712 |
| Autoglossed | 96,876 | 4,642 |
| Combined | 105,485 | 5,939 |

# Comparison with curated set



| Class | Example | C | GL (in C) | CB (in C) |
|---|---|---|---|---|
| *0001* | *tatuk* ‘have fever’ | 0 | 29 (1) | 11 (0) |
| *0010* | *tahai* ‘look for’ | 0 | 60 (4) | 24 (1) |
| *0011* | *luk* ‘bend’ | 25 | 4 (2) | 4 (2) |
| *0100* | *buk* ‘tie’ | 9 | 112 (10) | 50 (4) |
| *0101* | *tok* ‘drop’ | 0 | 3 (1) | 8 (1) |
| *0110* | *weel* ‘bathe’ | 1 | 2 (1) | 6 (0) |
| *0111* | *bel* ‘pull out’ | 10 | 1 (1) | 1 (0) |
| *1000* | *king* ‘long’ | 4 | 430 (38) | 247 (19) |
| *1001* | *mpang* think | 1 | 33 (8) | 35 (6) |
| *1010* | *aai* ‘add’ | 2 | 69 (17) | 97 (15) |
| *1011* | *kafia* ‘scratch’ | 184 | 25 (12) | 50 (6) |
| *1100* | *toq* ‘demolish’ | 19 | 59 (13) | 121 (18) |
| *1101* | *kolra* ‘cheat’ | 2 | 7 (1) | 35 (6) |
| *1110* | *momang* ‘clean’ | 3 | 13 (2) | 62 (8) |
| *1111* | *buuk* ‘drink’ | 75 | 7 (2) | 103 (27) |
|  | Total | 337 | 854 (113) | 854 (113) |

# Mismatch analysis



| Condition | Match | Mismatch | |
|---|---|---|---|
| | | Type 1 | Type 2 |
| Curated v. GL | | | |
| *I* | 72 | 20 | 21 |
| *II* | 90 | 17 | 6 |
| *III* | 50 | 61 | 2 |
| *IV* | 36 | 75 | 2 |
| Total | 248 | 173 | 31 |
| Curated v. CB | | | |
| *I* | **84** | 8 | 21 |
| *II* | 83 | 4 | 26 |
| *III* | **66** | 43 | 4 |
| *IV* | **54** | 56 | 3 |
| Total | **287** | 111 | 54 |

Type 1: Gaps in collected corpus

Type 2: Gaps in curated set

# Takeaways

- The system helped:
  - Identify mistakes in the IGT corpus
  - Hypothesize one previously unknown verb class
  - Hypothesize new members for known classes

- System-related issues:
  - Tagging nouns as verbs
    - Understandable in Abui!
  - Lack of phonological normalization