LREC-COLING 2024

EQ4ALL

# SignBLEU: Automatic Evaluation of Multi-channel Sign Language Translation

Jung-Ho Kim*, Mathew Huerta-Enochian*, Changyong Ko, and Du Hui Lee

EQ4ALL

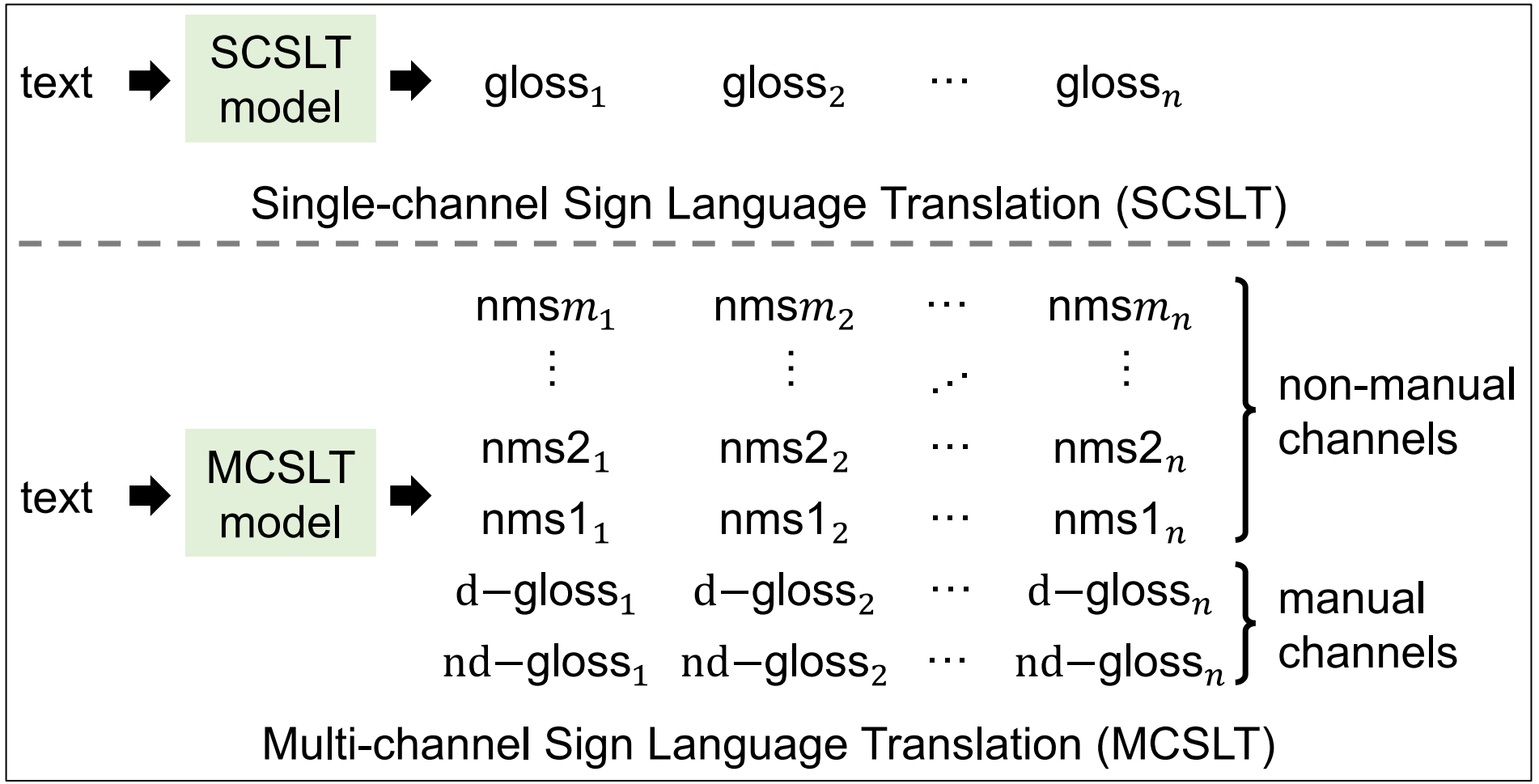
산업통상자원부
Ministry of Trade, Industry and Energy

IITP 정보통신기획평가원
Institute of Information & Communications Technology Planning & Evaluation

# Sign Language Translation (SLT)

- Sign Language:
  - Two common misunderstandings about sign languages:
    - Natural Language
    - Different from the regional spoken language
  - Can utilize the hands asynchronously, other body parts, space
- SLT
  - Translation between at least one sign language and another (usually spoken) language
  - Two directions:
    - Sign2Text - Text2Sign (this research is on Text2Sign)
    - Gloss-based, end-to-end, etc. Majority of T2S research has been on gloss-based
- Difficulty: Multiple articulators -> MCSLT

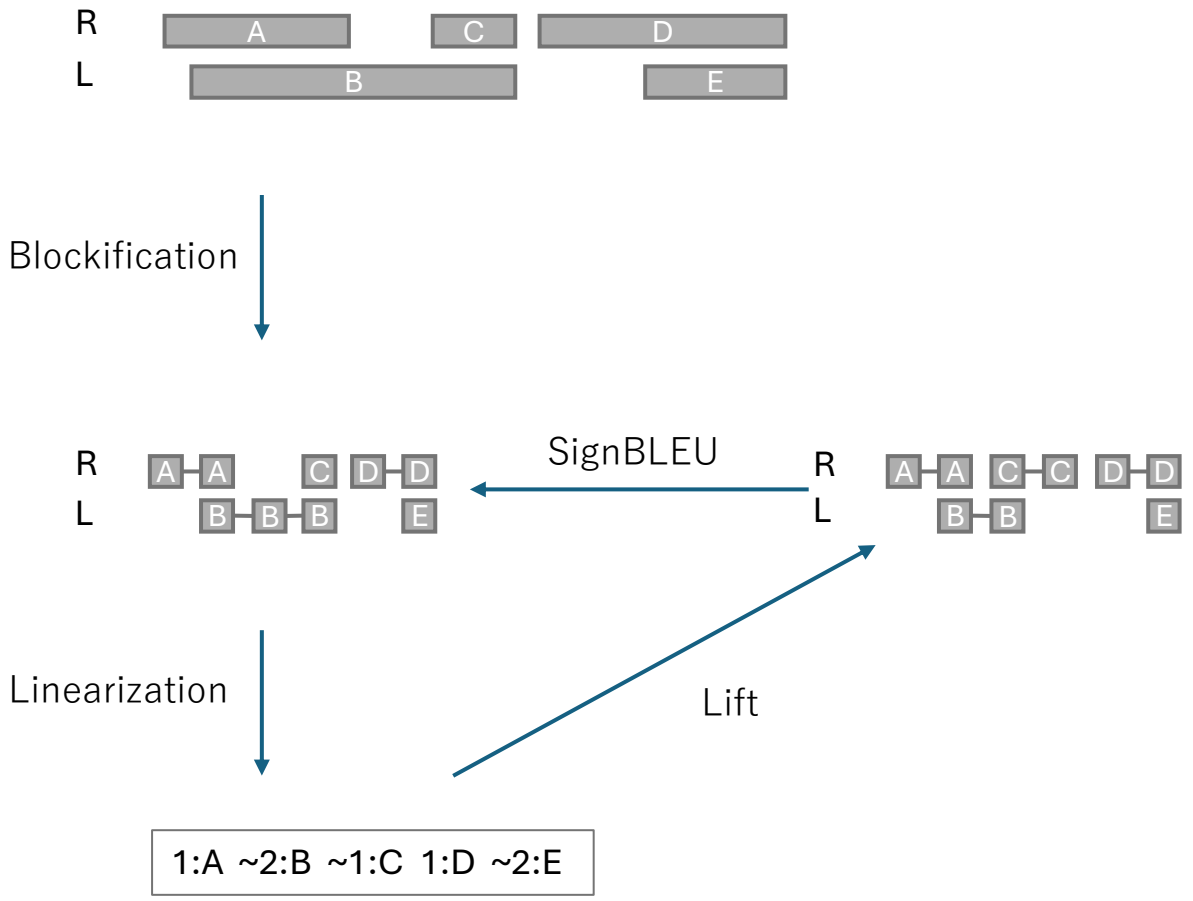
# Multi-Channel Sign Language Translation



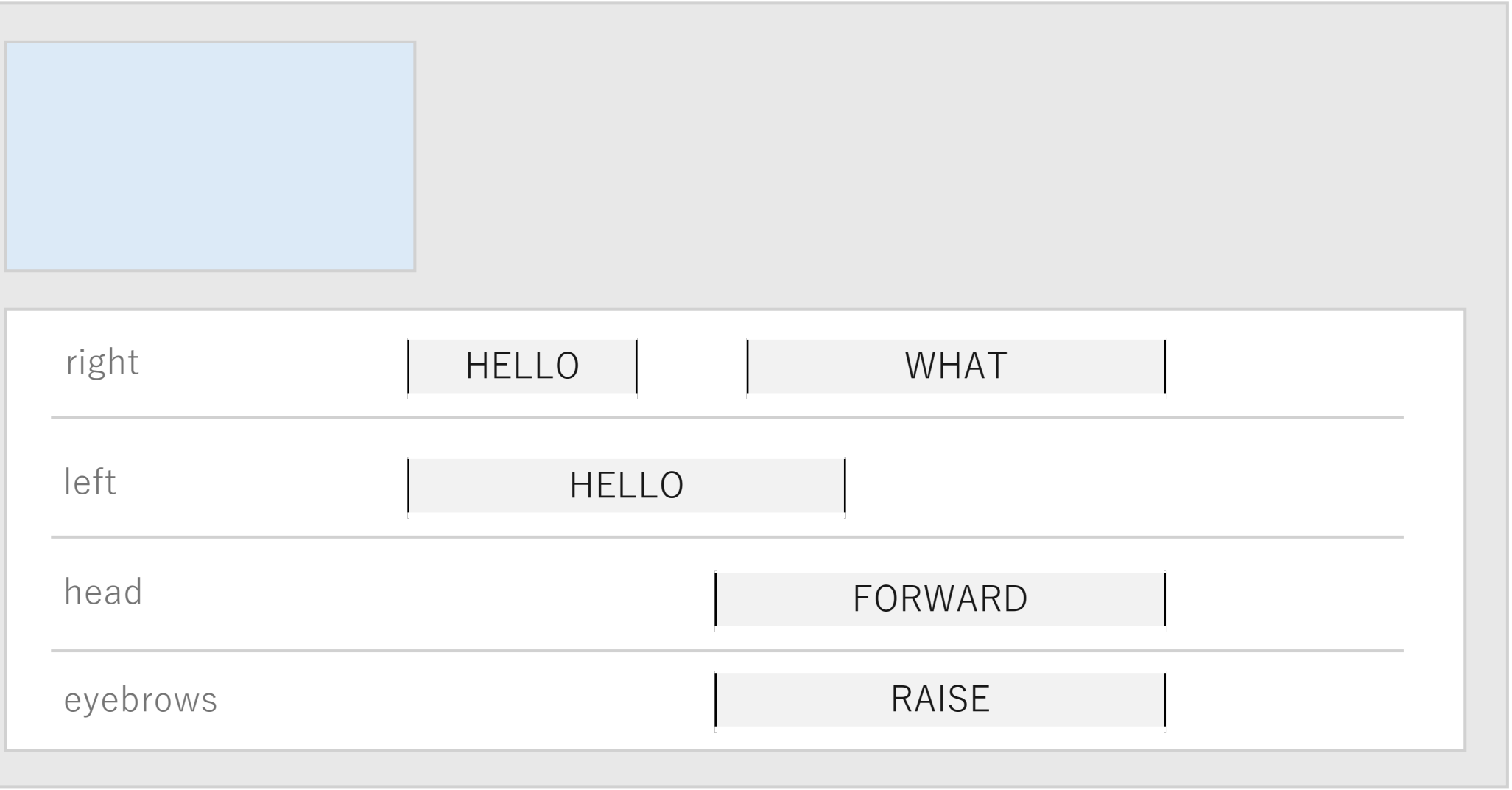

But this introduces a second problem:
what evaluation method is appropriate for multiple interacting, signing channels?

# Data Representation

- Three representations:
  - Raw annotation data
    - Time-aligned data
    - Multi-tier
  - Block (column) data
    - Overlap-aware
    - Duration-free
  - Linear Data
    - Partially overlap-aware
    - Duration-free

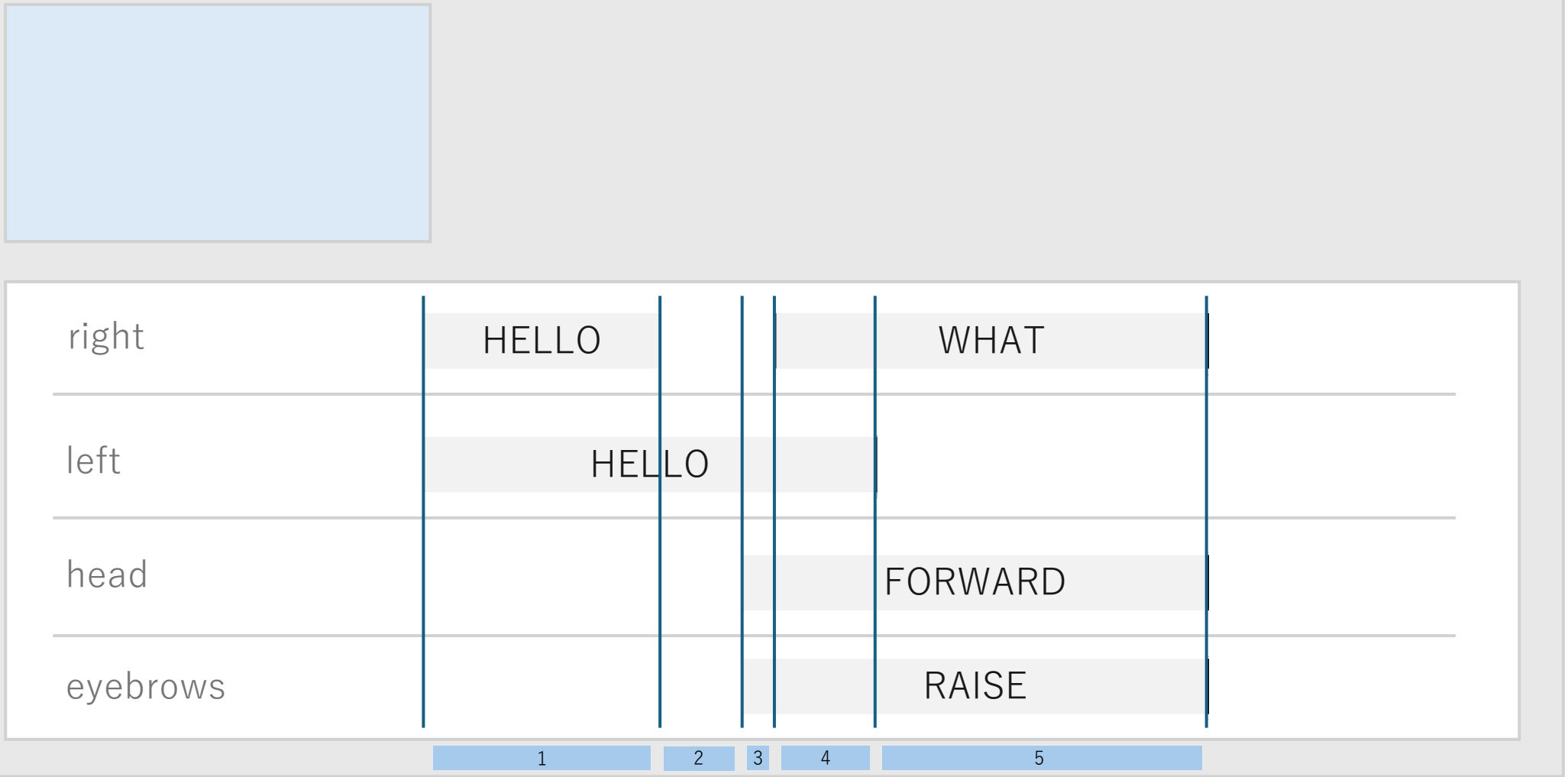


# Blockification: Raw data



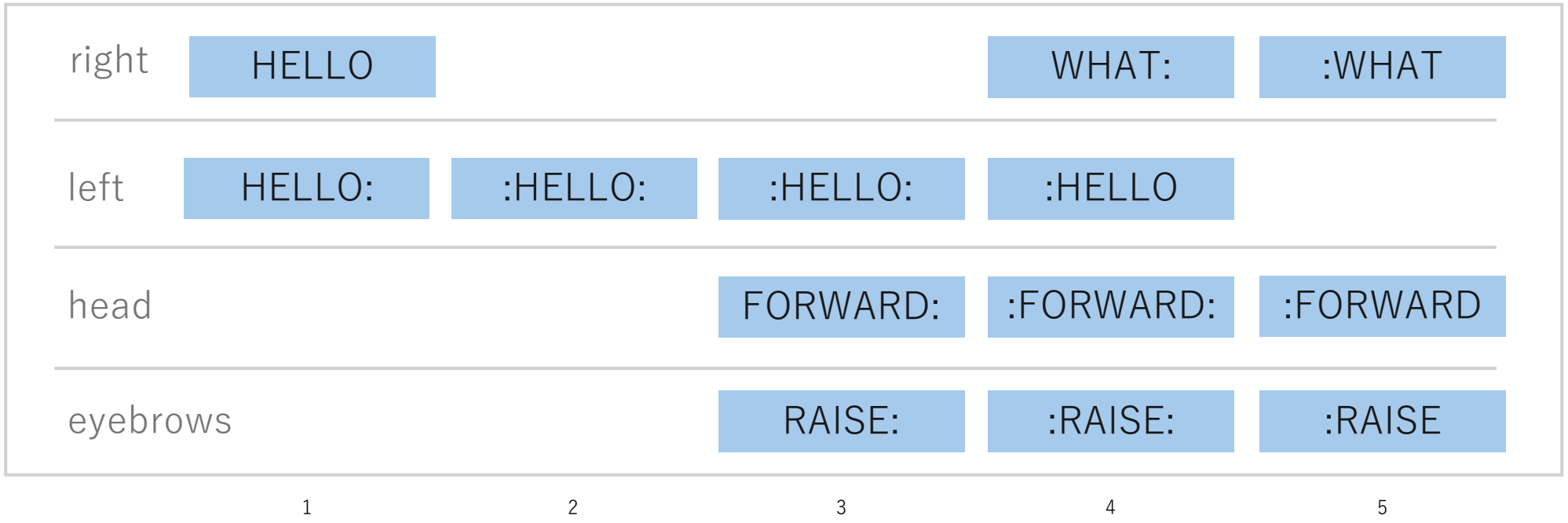

Raw annotation data

# Blockification: Segmented raw data




Segment on annotation start and end: 5 “blocks”

# Blockification: Sequential blocks




Remove duration information: 5 blocks representing gloss overlap

Continuation identifier “:”

# Blockification: JSON data


```
[
 {
  "right": "HELLO",
  "left": "HELLO:"
 },
 {
  "left": ":HELLO:"
 },
 {
  "left": ":HELLO:",
  "head": "FORWARD:",
  "eyebrows": "RAISE:"
 },
 {
  "right": "WHAT:",
  "left": ":HELLO",
  "head": ":FORWARD:",
  "eyebrows": ":RAISE:"
 },
 {
  "right": ":WHAT",
  "head": ":FORWARD",
  "eyebrows": ":RAISE"
 }
]
```


Easily expressed as JSON data.

Format places emphasis on gloss overlap instead of duration to focus on multi-articulator expressions.

# BLEU - Review existing metrics in NLP

EQ4ALL

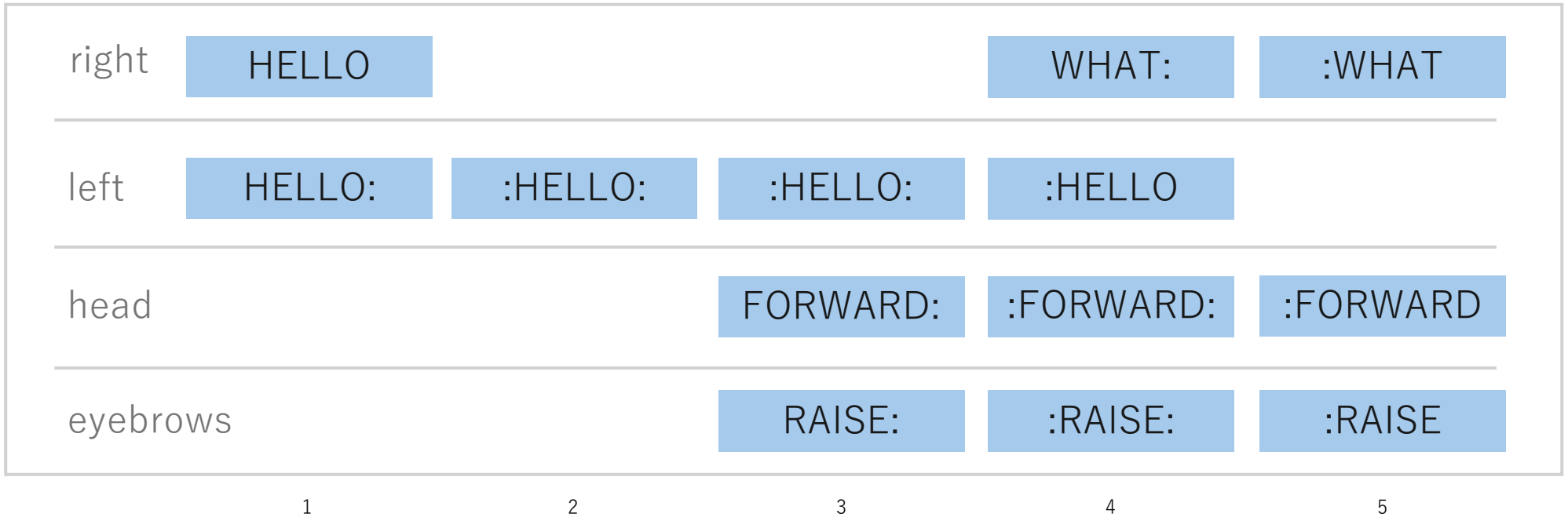
- Current most common method for translation eval is BLEU
- BLEU is an n-gram-based modified precision score
  - Precision(n) = precision(ngram(candidate), ngram(references))
  - BLEU-N = mean(Precision(1), ⋯, Precision(N)) * Penalty

- Goal: Adapt BLEU to MCSLT

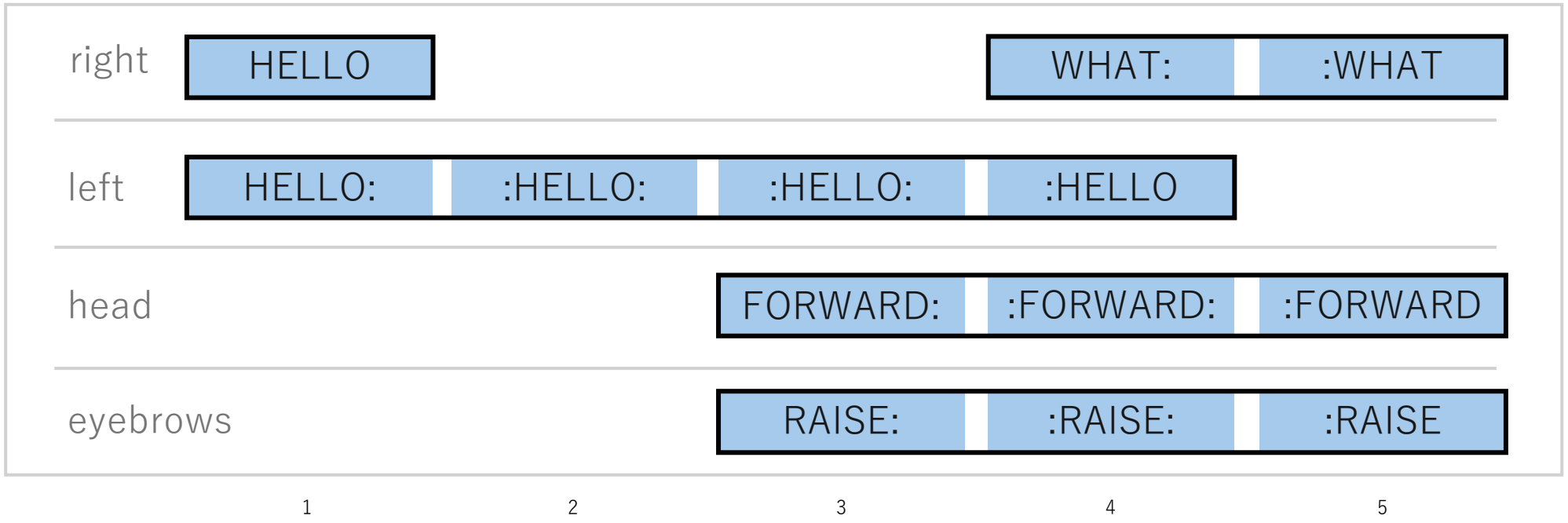
# SignBLEU

- Essentially BLEU, but with modified n-grams
  - Temporal Grams: traditional n-grams along each channel
    - “t1-grams”, “t2-grams”, “t3-grams”, “t4-grams”
  - Channel Grams: subsets of glosses (of size n) in each block.
    - “c1-grams”, “c2-grams”, “c3-grams”, “c4-grams”

# Temporal and Channel Grams

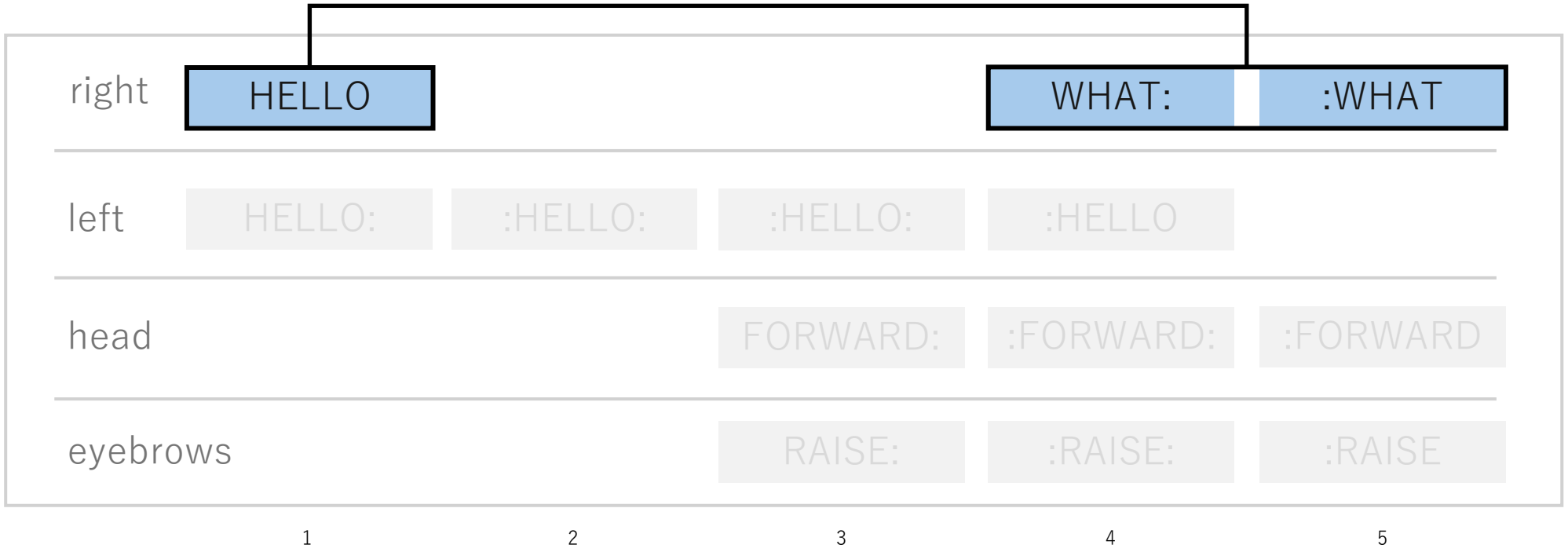


# T1-Grams



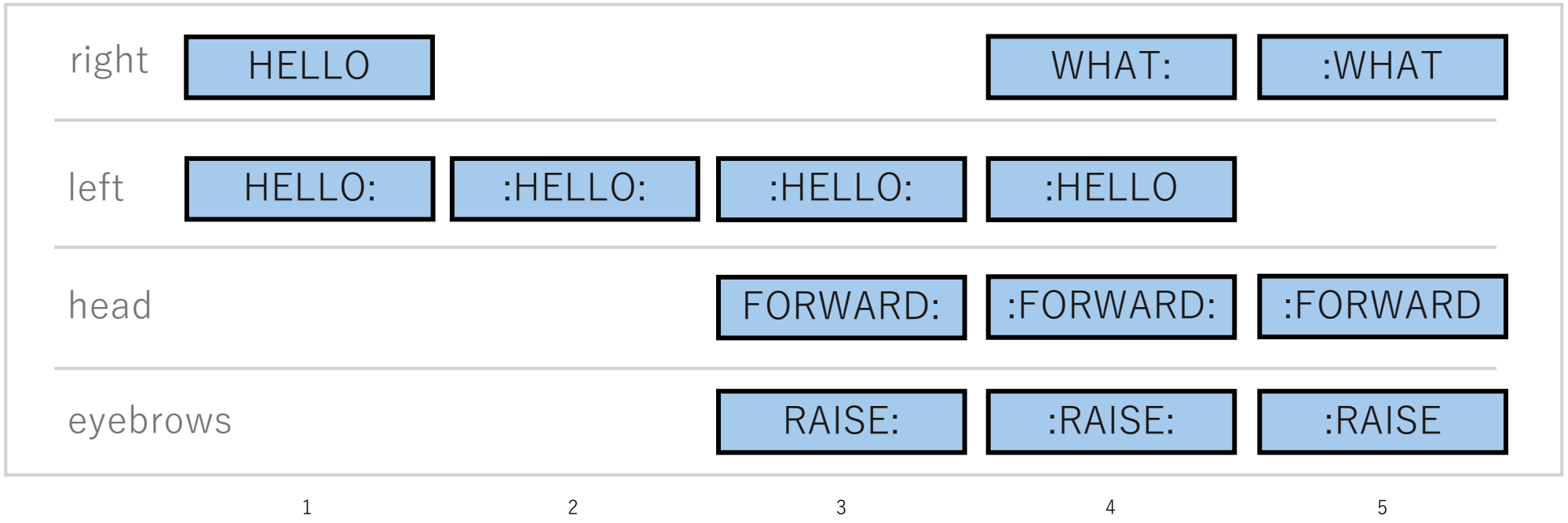

A total of 5 t1-grams

# T2-Grams




Only 1 t2-gram.

# C1-Grams

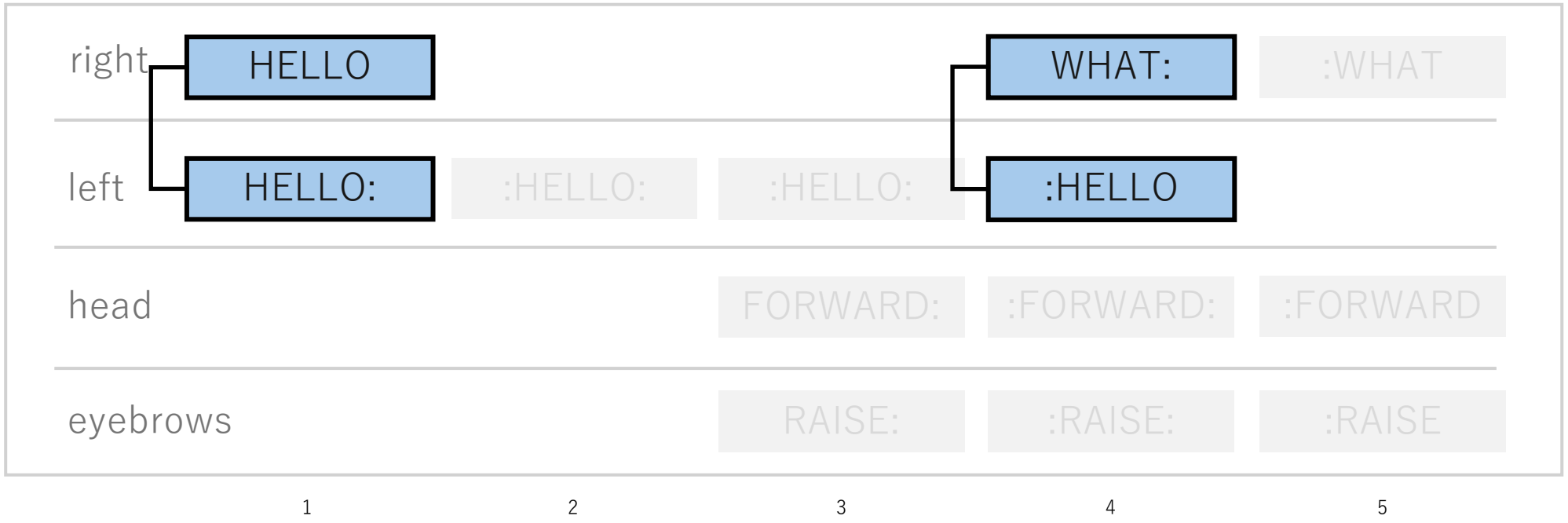



A total of 13 c1-grams.

The default SignBLEU does not use c1-grams:

- Overlap with t1-grams.
- Large set.

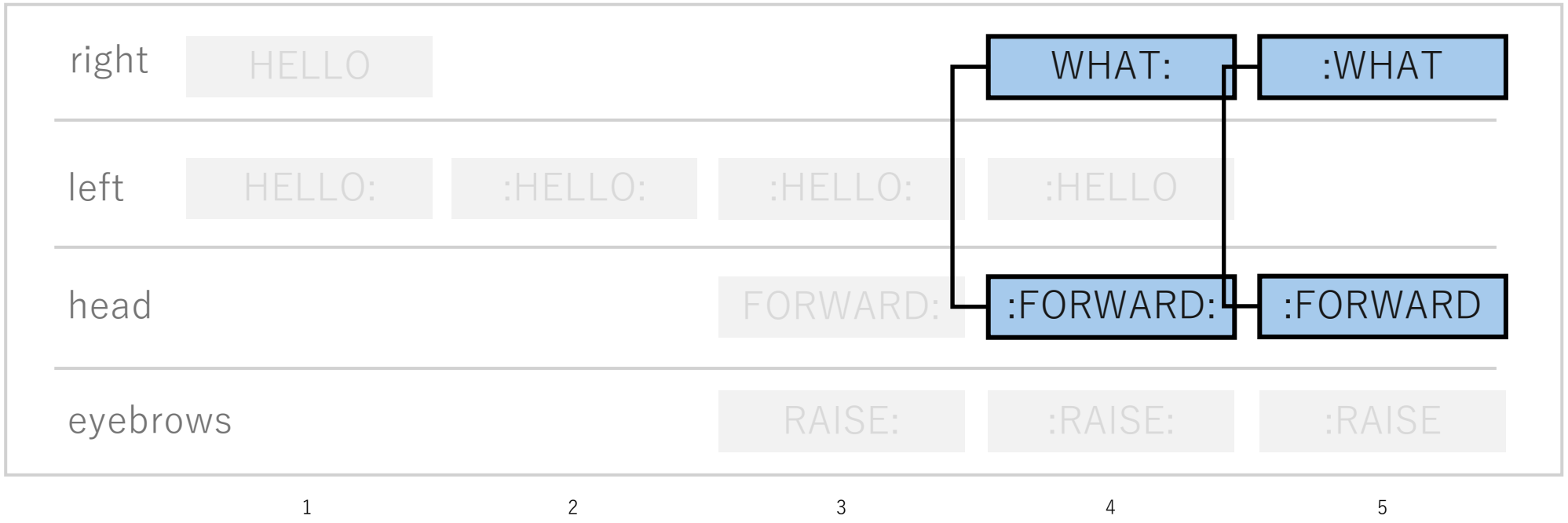
# C2-Grams




C2-grams spanning the left and right channels.

(A total of 13 c2-grams.)

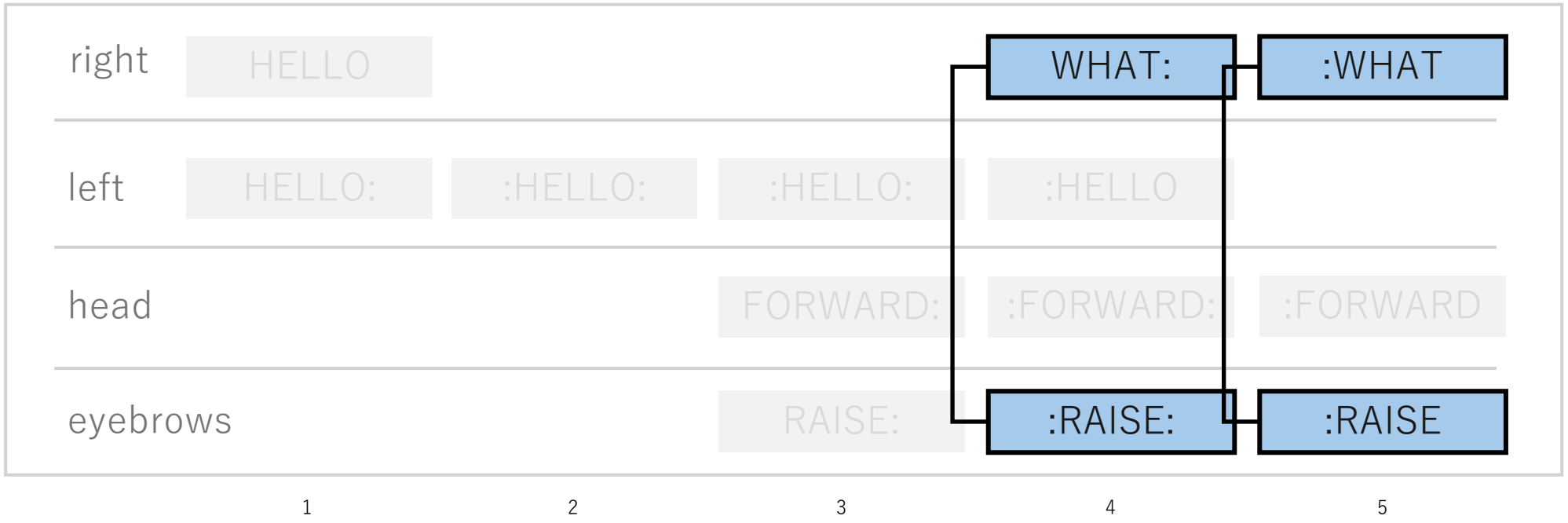
# Channel Grams: Example c1, c2 grams




C2-grams spanning the right and head channels.

(A total of 13 c2-grams.)

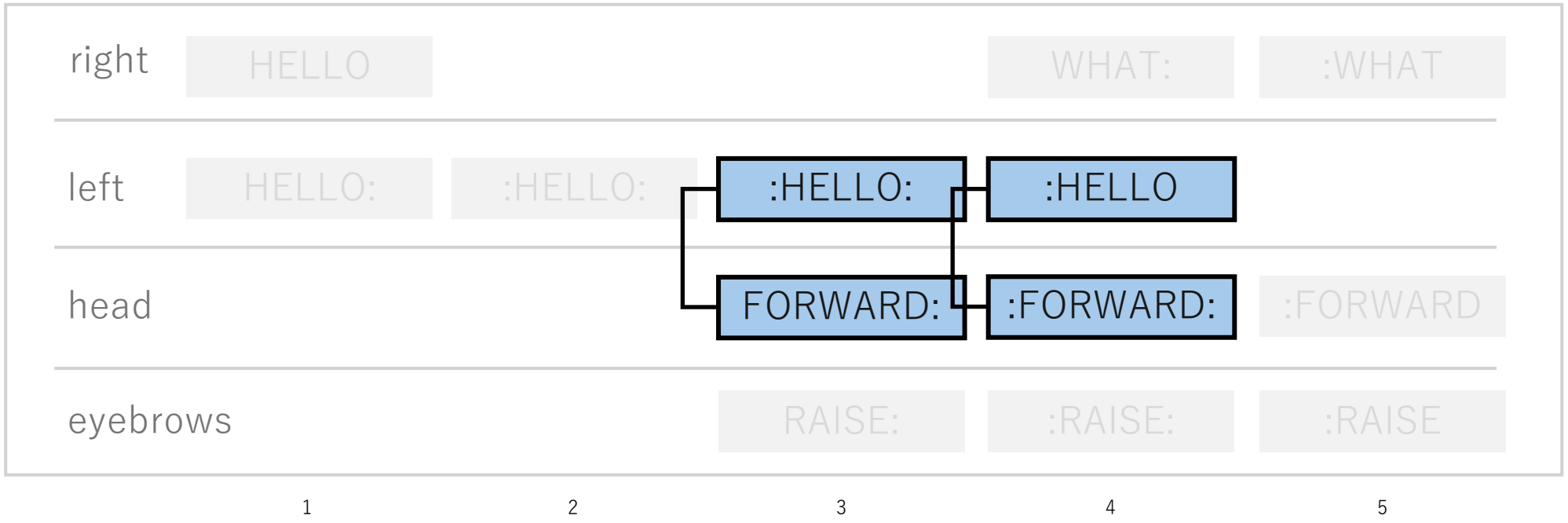
# Channel Grams: Example c1, c2 grams




C2-grams spanning the right and eyebrow channels.

(A total of 13 c2-grams.)

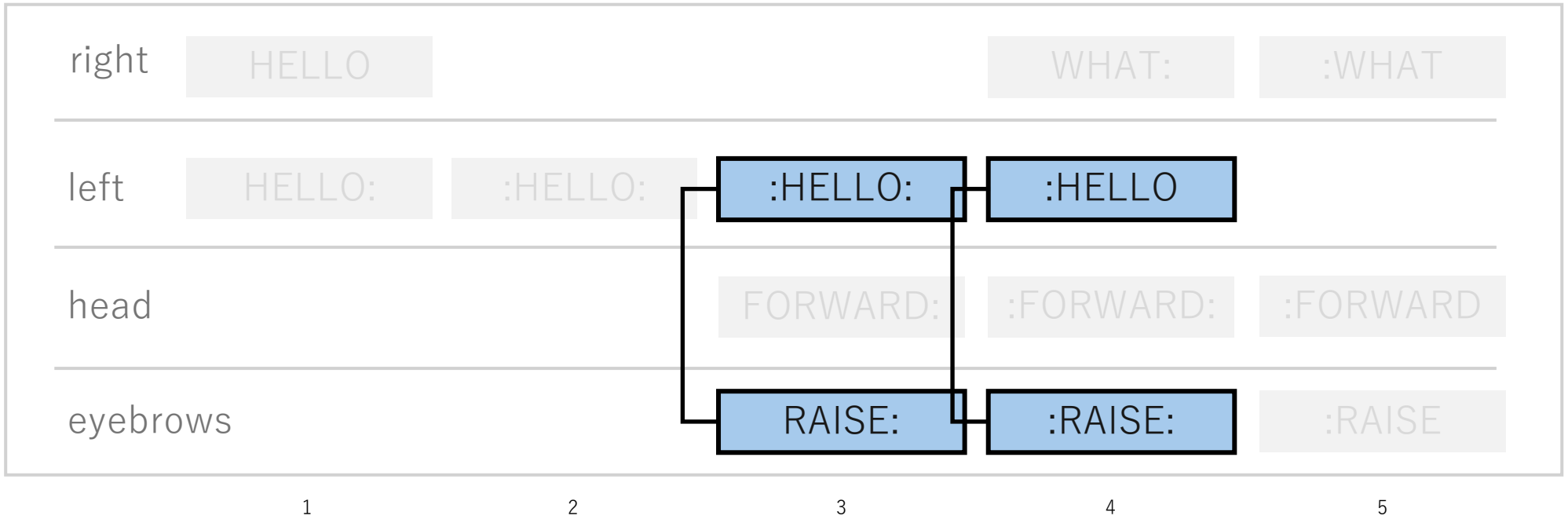
# Channel Grams: Example c1, c2 grams




C2-grams spanning the left and head channels.

(A total of 13 c2-grams.)

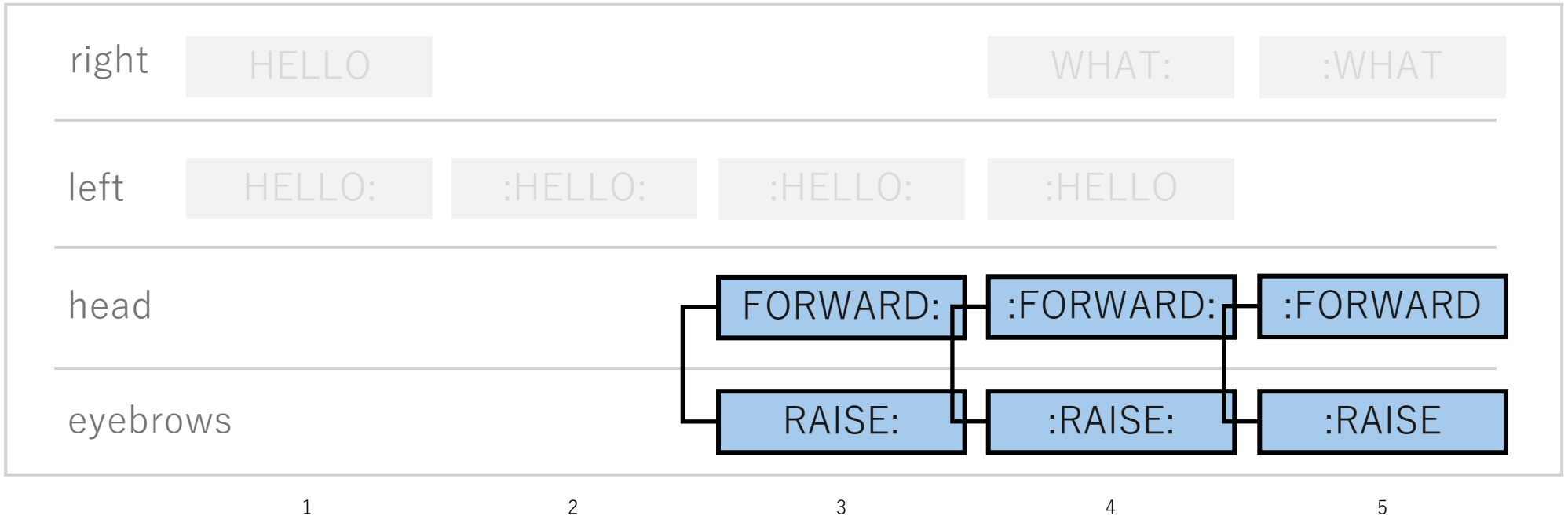
# Channel Grams: Example c1, c2 grams




C2-grams spanning the left and eyebrow channels.

(A total of 13 c2-grams.)

# Channel Grams: Example c1, c2 grams




C2-grams spanning the head and eyebrow channels.

(A total of 13 c2-grams.)

# SignBLEU

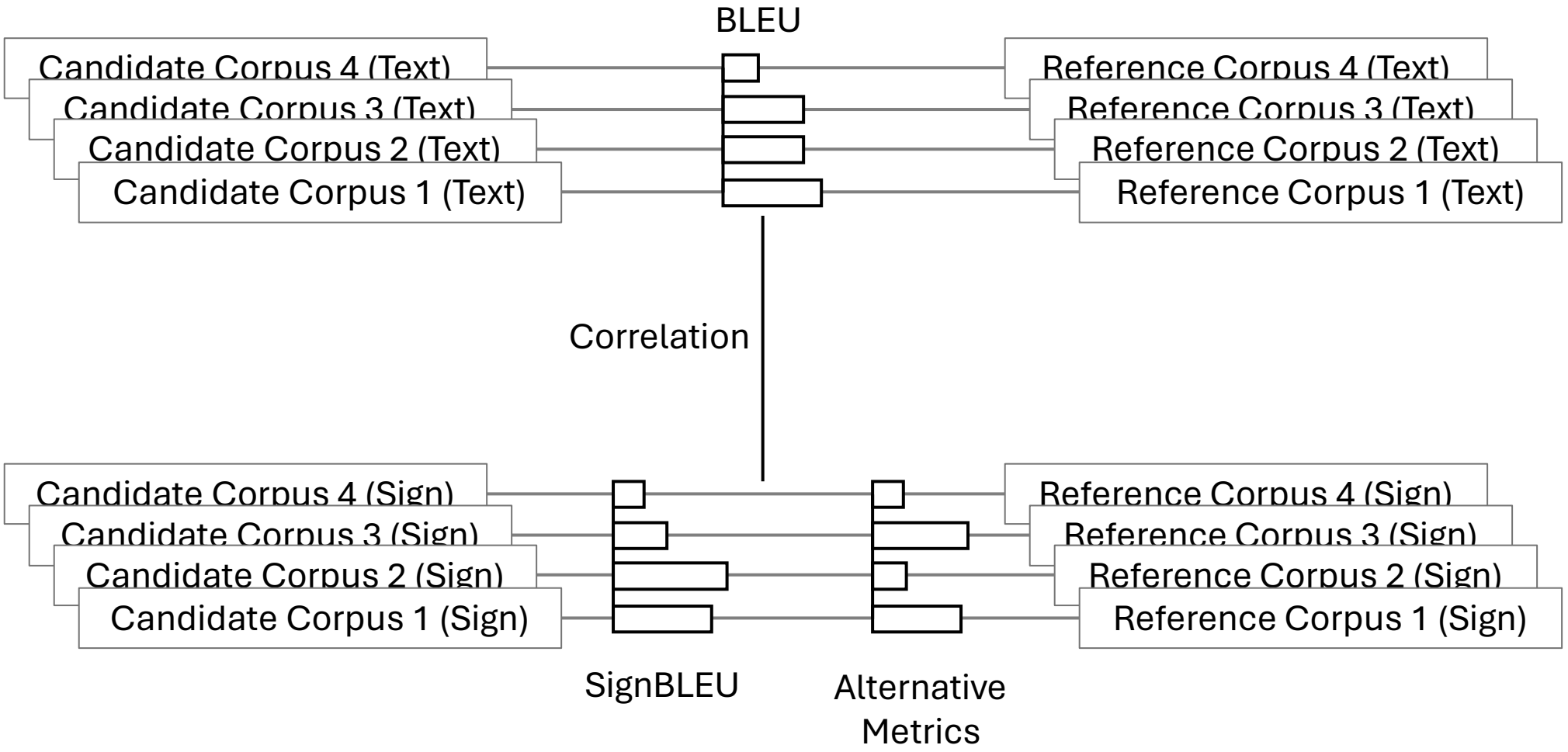
- Uses modified n-grams.
  - Temporal Grams: traditional n-grams along each channel
  - Channel Grams: subsets of glosses (of size n) in each block
  - 2D
    - Poor performance
    - High time complexity
- All-channel and Manual SignBLEU variants
- Algorithm equivalent to BLEU with t-grams and c-grams
  - SignBLEU-t2c3 => SignBLEU using t1, t2, c1, c2, c3 grams.

# Experimental Results I

- System-Level: Text-Side BLEU Correlation
  - Three datasets:
    - The Public DGS Corpus (PDC), NIASL2021 (NS21), and Boston University's The National Center for Sign Language and Gesture Resources corpus (NCSLGR).
    - Selected for variation in both language structure and annotation methods.

# Experimental Results I

- System-Level: Text-Side BLEU Correlation

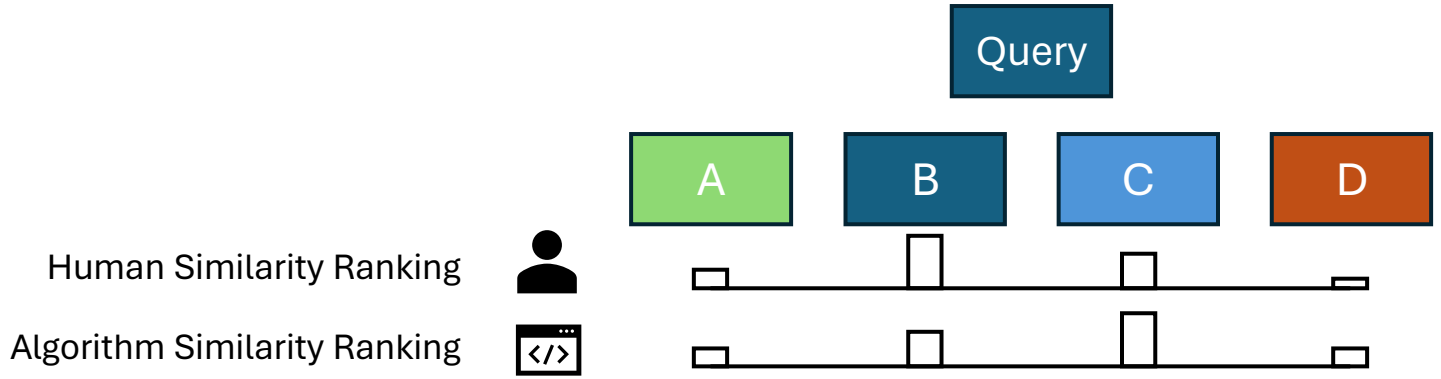


# Experimental Results I

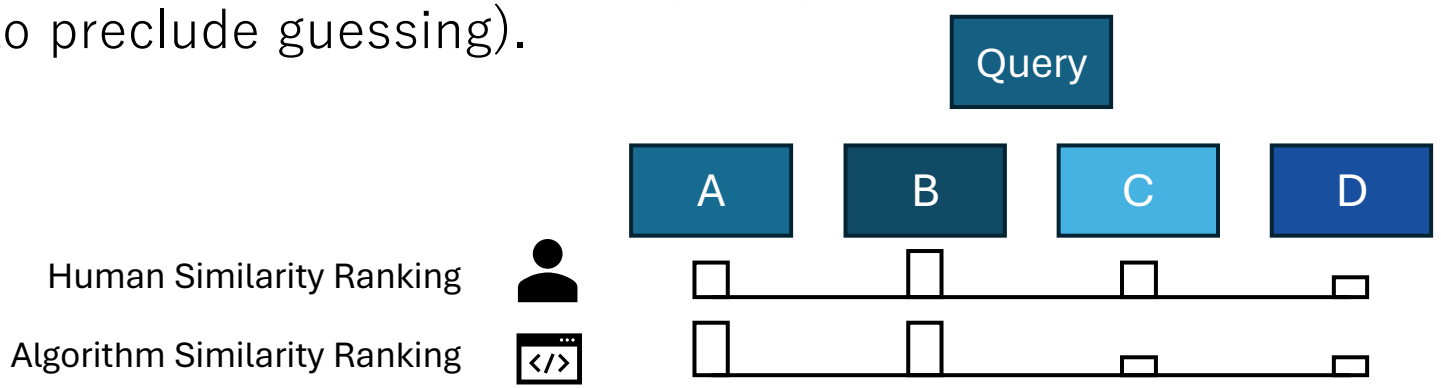
- System-Level: Text-Side BLEU Correlation
  - Three datasets:
    - The Public DGS Corpus (PDC), NIASL2021 (NS21), and Boston University's The National Center for Sign Language and Gesture Resources corpus (NCSLGR).
    - Selected for variation in both language structure and annotation methods.
  - Results
    - Manual SB-t2c1 or t3c1 scored highest.
      - SB-txcy scored in top 2-3 for PDC and NCSLGR.
      - Manual SB is sufficient for many NS21 cases.
    - Ideal parameters are dependent on the dataset. This makes sense—PDC does not use many two-handed annotations and only has one non-manual channel. This will differ from NCSLGR which includes a wide variety of non manual annotations.

# Experimental Results II

- Segment-Level: Sign Video Similarity Ranking
  - Two scenarios
    - Easy: Videos cover different topics or use different vocabulary




    - Hard: Videos cover same topic, single signer with same clothes or multi-signer(to preclude guessing).

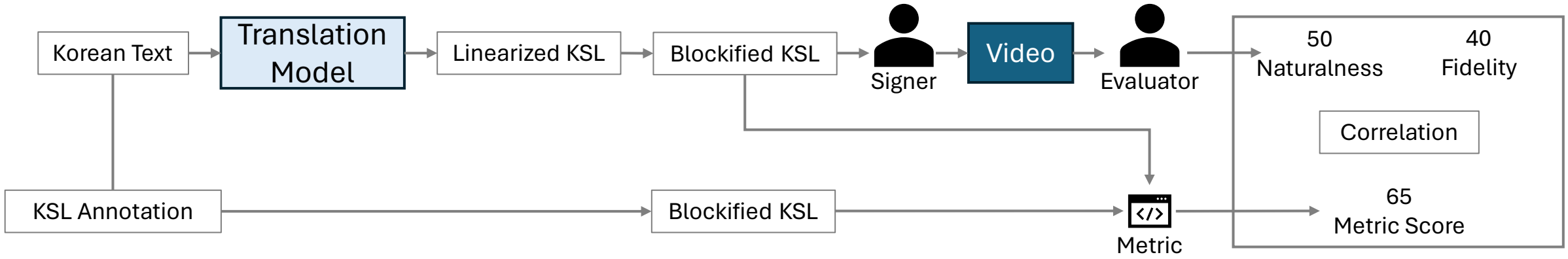


# Experimental Results II

- Segment-Level: Sign Video Similarity
  - Two data splits
    - Easy: Videos cover different topics or use different vocabulary
    - Hard: Videos cover same topic, single signer with same clothes or multi-signer(to preclude guessing).
    - Results:
      - Easy: Manual BLEU-2 highest correlation, but little difference between BLEU and SignBLEU scores.
      - Hard: SignBLEU-t1c2 highest correlation.
    - This makes sense.
      - When signing content is different, a simplistic view of the multiple channels should be sufficient for similarity analysis.
      - When signing content is similar, better analysis of the multi-channel data (including non-manuals) is required.

# Experimental Results III

- Segment-Level: Translation Naturalness and Fidelity

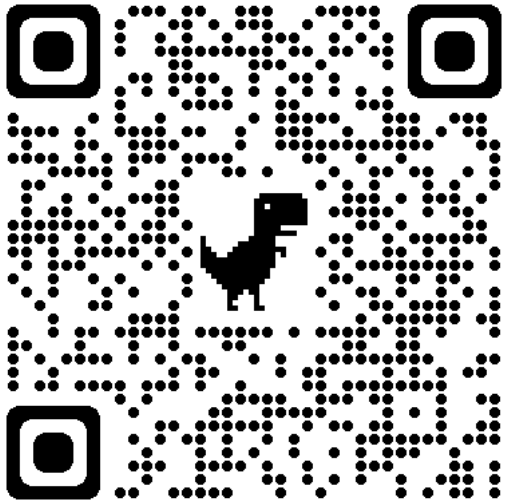


# Experimental Results III

- Segment-Level: Translation Naturalness and Fidelity
  - Naturalness:
    - Manual SB-t1c2 best.
      - Overlapping manual glosses were the most important factor in translation naturalness.
      - Note: This is for our specific test conditions, not commentary on the importance of non-manuals.
    - Many of the standard NLP metrics applied to linearized translations (using both manual and all-channel variants) showed negative correlation. This is further evidence that they are not appropriate for MCSLT.
  - Fidelity:
    - Manual SB-t3c2 best (with manual SB-t(1-2)c(1-2) all similarly high). Again, based on the SignBLEU algorithm, manual annotations were sufficient to evaluate fidelity.

# Takeaways

- SignBLEU requires some parameter tuning based on the characteristics of the target data for best results.
- We released the SignBLEU code at: https://github.com/eq4all-projects/SignBLEU



- We hope to work with the community to improve SignBLEU.