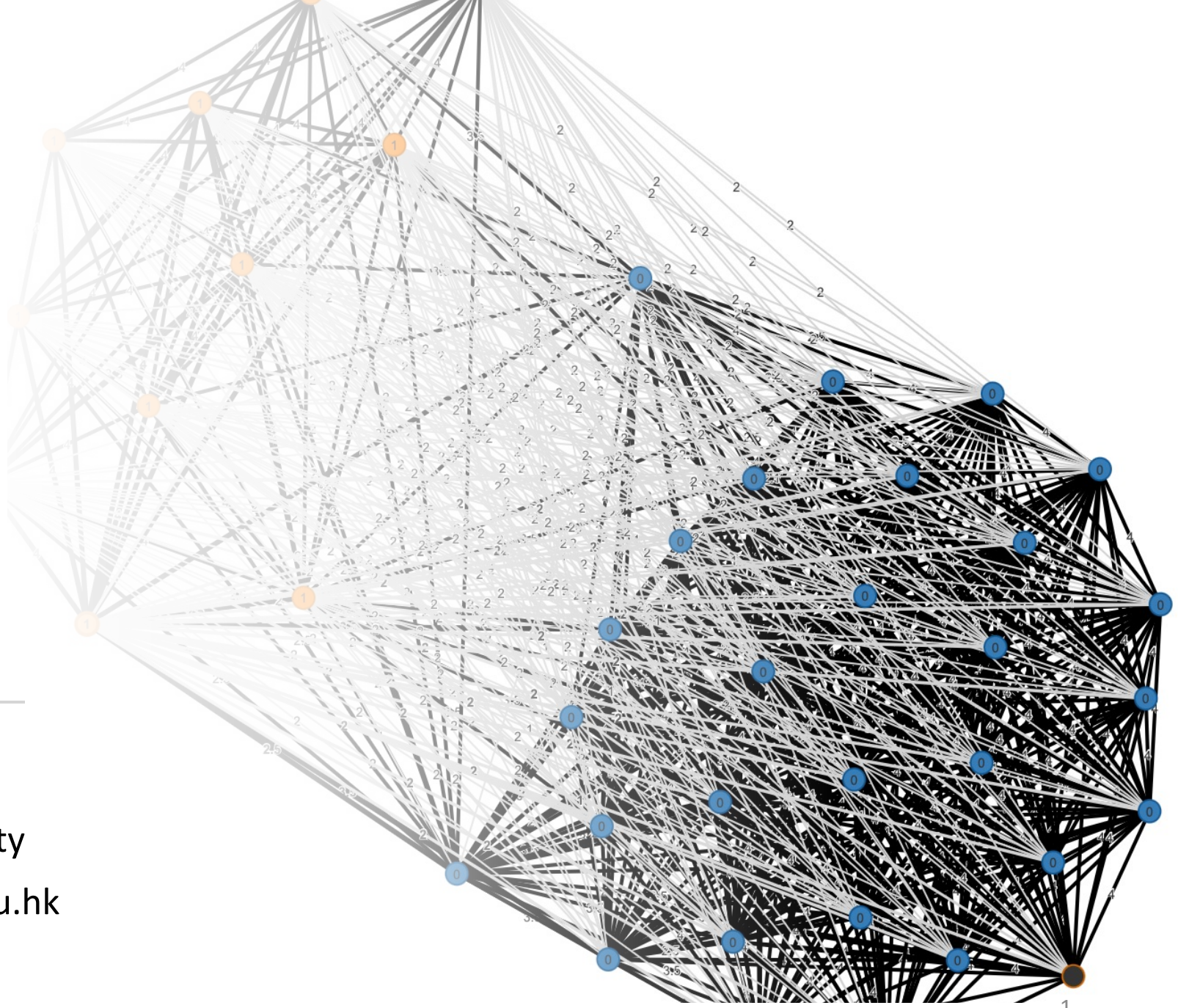
DWUG Meets Chinese: Visualizing Chinese Semantic Shifts with Expert Judgments

Jing Chen

The Hong Kong Polytechnic University

Contact: jing95.chen@connect.polyu.hk



Reference: Chen, J., Chersoni, E., Schlechtweg, D., Prokic, J., & Huang, C.-R. (2023). ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection. In N. Tahmasebi, S. Montariol, H. Dubossarsky, A. Kutuzov, S. Hengchen, D. Alfter, F. Periti, & P. Cassotti (Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change* (pp. 93–99). Association for *Computational Linguistics (ACL).*

Outline

- Background
- Methodology
  - Data Source and Selection
  - Human Annotation
- Graph Representation
- Quantifying Changes: Metrics for Semantic Change
  - Binary Change
  - COMPARE
  - Graded Change(JSD)
- Conclusion and Limitation

Background

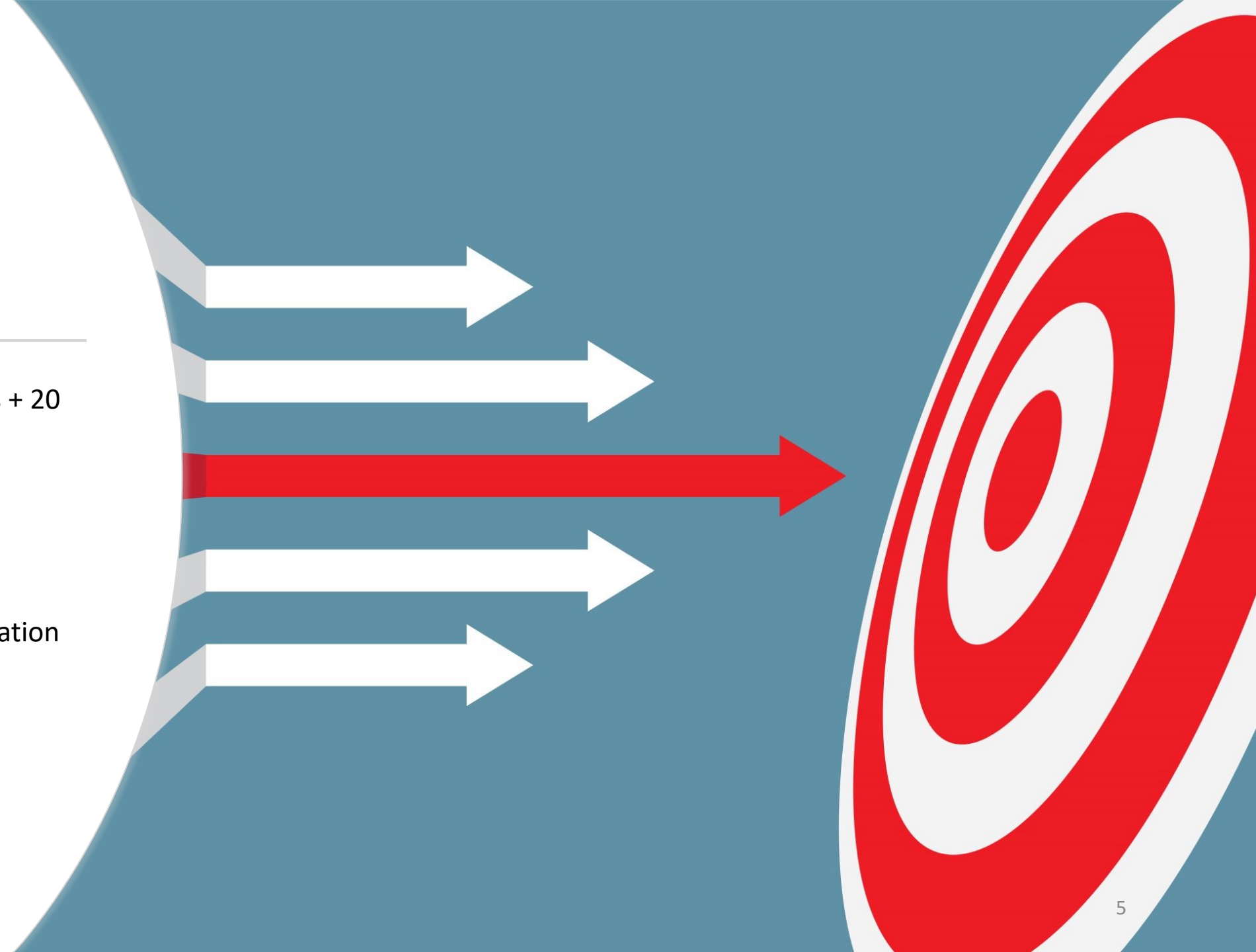
- Language models is now reshaping research paradigms
- LSCD: Lexical Semantic Change Detection
- Benchmarks
  - DURel and DWUG
- DWUG Meets Chinese? ChiWUG is coming!

Data

- Newspaper Archive: People’s Daily
- Historical Phases: the transformative phase of China's Reform and Opening Up
  - Pre-: 1954 -- 1978
  - Post: 1979 -- 2003

Target words

- 40 targets: 20 changed ones + 20 fillers
- Selection criteria
- Filtering mechanism
- Frequency balance
- Manageable scale for annotation



Targets	Sentences	Pairs	Avg Tokens per Sent.
40	1600	31,200	53.39

Table 2: Statistics of usage. *Avg Tokens per Sent.* refers to the average number of characters in sampled sentences

Usage pairs

- Forty sentences per period were randomly sampled from the dataset for each target word
- Each target word is represented by two sets of 20 sentences each, from earlier and later periods

Human Annotation

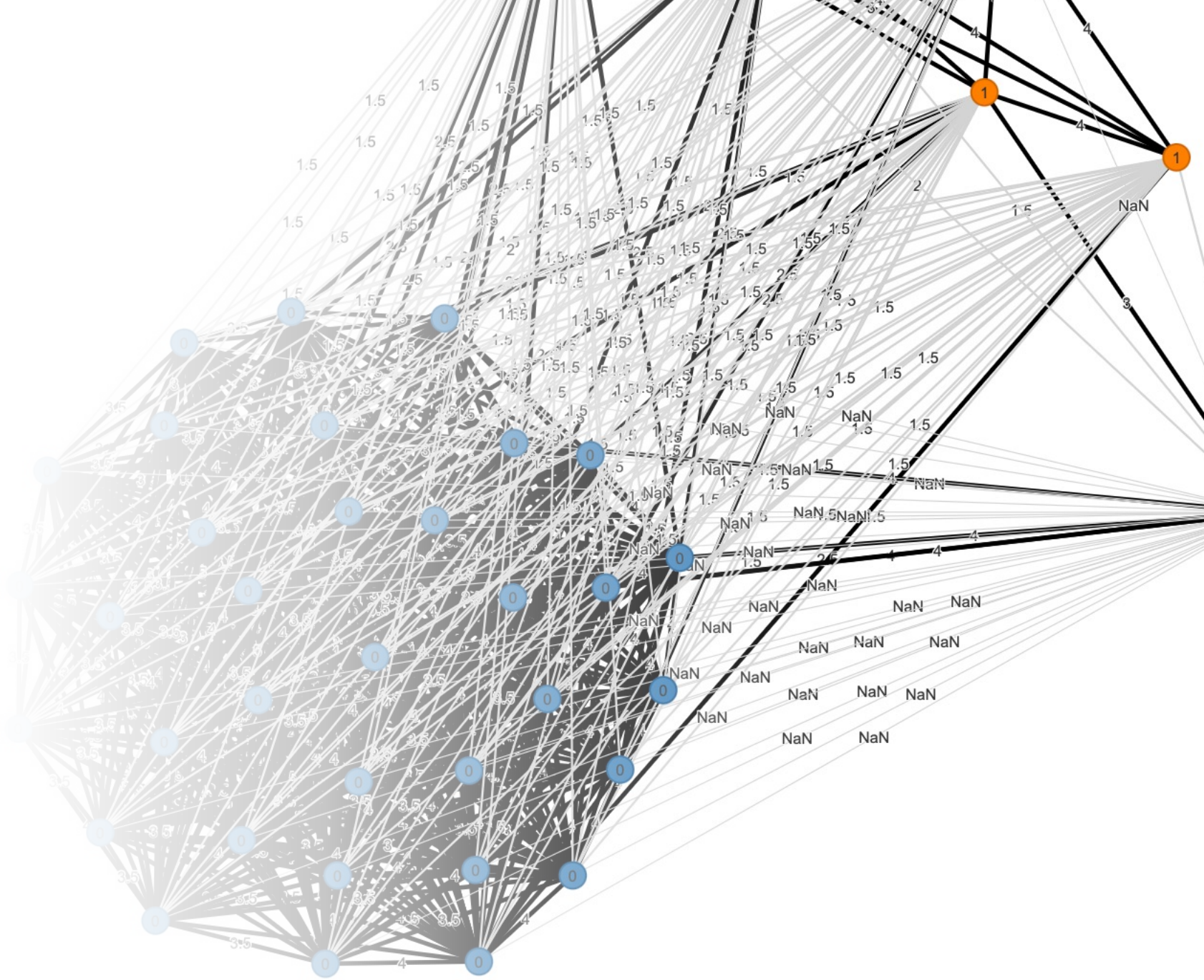
- 61K human judgments
- 4 native speakers, graduate students majored in Chinese Linguistics
- Semantic proximity: 1-4
- Annotation load: 2 annotator take a half consisting of 10 changed words and 10 fillers(random sampled)
- High inter-rater agreement: 0.691 for spearman, 0.602 for Krip-pendorff's alpha

Periods	n	N/V/A	|U|	AN	JUD	AV	SPR	K
1954-2003	40	10/22/8	1,599	4	61k	2	.691	.602

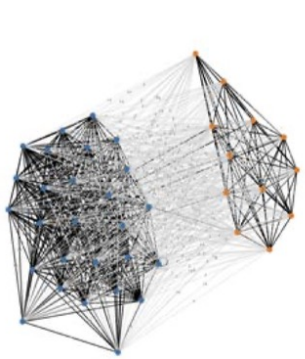
Table 3: Statistics of target words in ChSemShift. n = the number of usages, $N/V/A$ = the number of nouns, verbs and adjectives, $|U|$ = the total number of usages. One usage pair was discarded during the annotation due to the context ambiguity. AN = the number of annotators, JUD = the number of judgments, AV = the average number of annotations per usage pair, SPR = weighted mean of pairwise Spearman score, K = Krippendorff's alpha.

Graph Representations

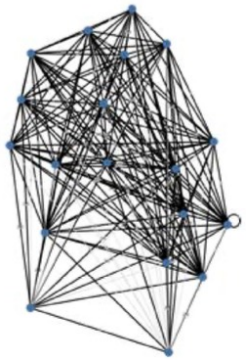
- Aggregation: correlation clustering (Bansal et al., 2004)
  - usage pairs with scores 3 and 4 as the same sense
  - while scores 1 and 2 were considered as different senses



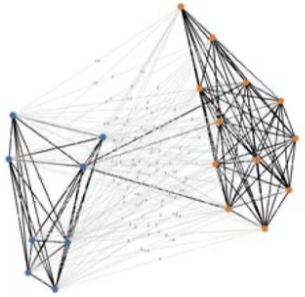
Word Usage Graphs of 下海 xiahai, “go into the sea; to venture”



(a) Full graph

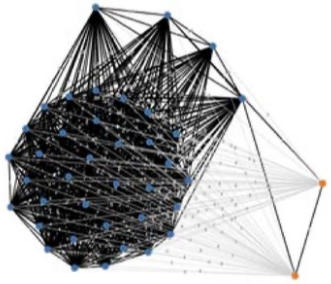


(b) Subgraph for the first period

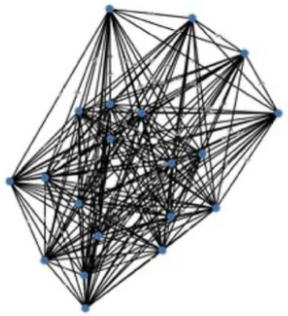


(c) Subgraph for the second period

Word Usage Graphs of 病毒 bingdu, ‘computer virus’ and “viral infection”



(a) Full graph



(b) Subgraph for the first period 2

(c) Subgraph for the second period

$$B(w) = \begin{cases} 1 & \text{if for some } i, D_i \leq k \text{ and } E_i \geq n, \\ & \text{or vice versa.} \\ 0 & \text{otherwise} \end{cases}$$

$$C(W_{1,2}) = \frac{1}{|W_{1,2}|} \sum_{x \in W_{1,2}} x \qquad (1)$$

$$JSD(P,Q) = \sqrt{\frac{KLD(P\|M) + KLD(Q\|M)}{2}} \qquad (2)$$

where:

$$KLD(P\|Q) = \sum_{i}^{K} log_2(\frac{p_i}{q_i}), \quad M = \frac{(P+Q)}{2}$$

Quantifying Changes: Metrics for Semantic change

- 3 metrics
  - 1 for binary change
  - 2 for graded change
    - COMPARE: (1)
    - Jensen-Shannon Distance: (2)

- Strong correlation between graded change and COMPARE metric.
- Correlation of both scores with binary change.
- Instances like "软 (ruan)" demonstrate significant graded change without binary change.
- Binary change in words is associated with varying degrees of graded change.

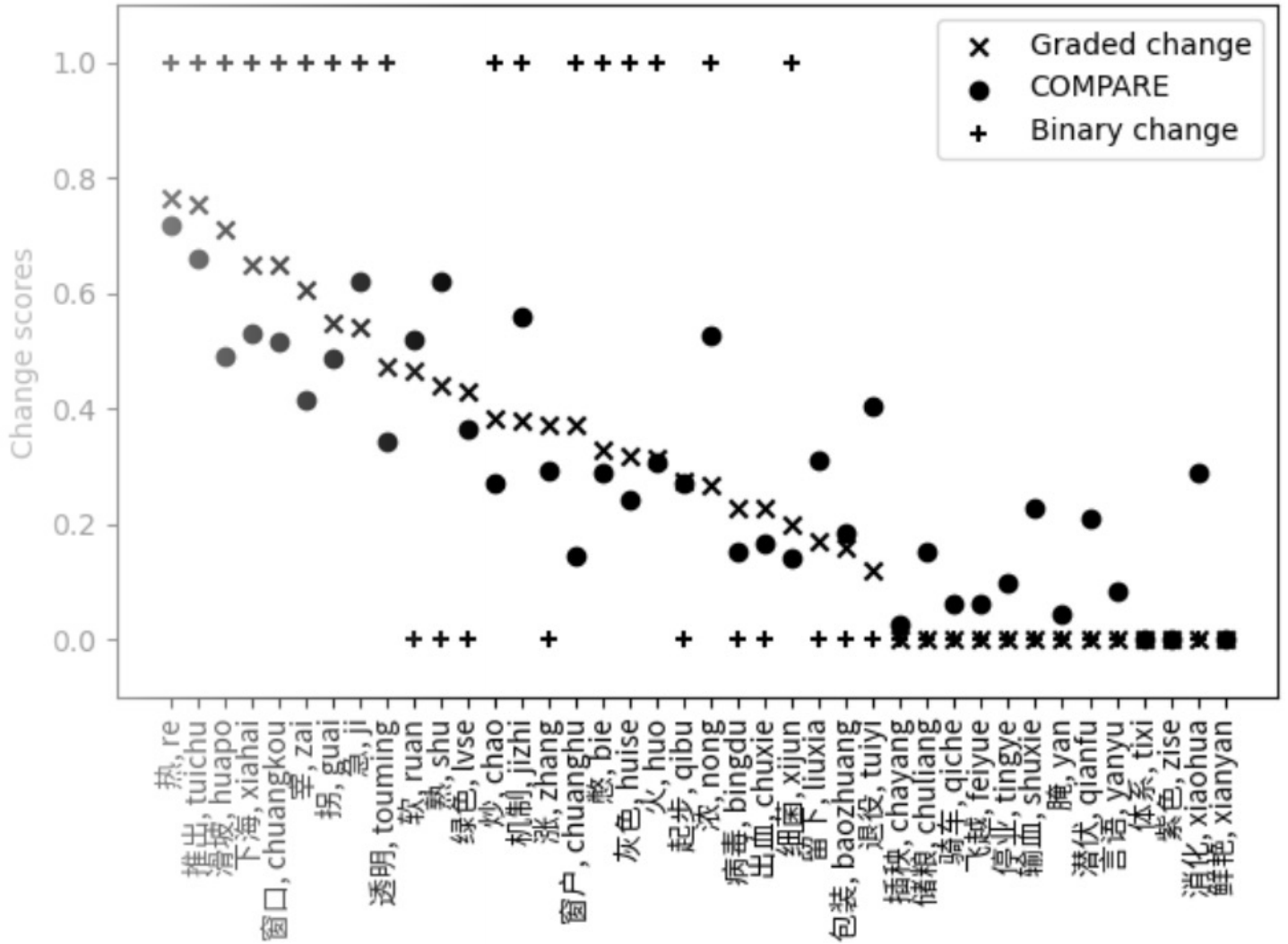


Figure 4: Change scores inferred on the WUGs resulting from our annotation. The COMPARE score was ma
with $f(x) = 1 - \frac{1}{3}(x - 1)$ to fit the range of the other scores and to follow their direction (higher values mean
change).

Conclusion and Limitation

This study presents the first graph-based evaluation dataset for Chinese LSCD in the context of the Reform and Opening-up period.

It populates 40 word usage graphs based on over 61k human judgments and has high inter-rater agreement.

This study investigated the period from the 1950s to the 2000s, based on a regional newspaper dataset, may only partially reflect the broader linguistic changes.