

Coverage Is Not Enough: A Controlled Study of Annotation Selection for Low-Resource Sign Language

Gyanendra Shrestha
Florida State University
gshrestha@fsu.edu

Liting Zhang
University of West Florida
lzhang@uwf.edu

Abstract

Expert annotation is a central bottleneck in sign-language corpus construction. A natural assumption is that annotation budgets should prioritize diverse or underrepresented signing patterns. We test this in a controlled retrospective study on PHOENIX14T (German Sign Language) and Isharah (Saudi Sign Language), across gloss recognition, translation, multiple budgets, and three seeds, comparing random sampling against geometric, active-learning, pose-derived articulatory, and privileged selectors. Articulatory coverage is generally weak across all four corpus–task settings and rarely outperforms random sampling. A density analysis shows it selects locally atypical clips even after controlling for duration; removing the sparsest 10–20% helps at larger budgets but never enough to beat random. A privileged gloss-diversity oracle also performs poorly despite selecting comparatively typical clips, showing that atypicality is not the whole explanation. Facility-location representativeness is the strongest feasible selector in three of four largest-budget settings, with lower WER than random in 16 of 18 seed-matched recognition comparisons, compared with only 2 of 18 for articulatory-only selection. The results suggest pose-derived articulatory coverage and gloss-type coverage, as operationalized here, are unreliable proxies for annotation utility in low-resource sign-language corpus construction. The source code is available at <https://github.com/gyan-shrestha/WSLP-2026>.

1 Introduction

Continuous sign-language corpora are expensive to annotate because glossing, translation, segmentation, and temporal alignment require specialist expertise. Recorded video can therefore substantially exceed manually annotated material: Content4All released approximately 190 hours of broadcast signing, but only about 20 hours received man-

ual alignment by Deaf experts and interpreters (Camgöz et al., 2021). This bottleneck is particularly consequential for underrepresented sign languages (De Sisto et al., 2022).

A corpus team with a fixed budget must decide which clips to annotate before their linguistic content can be observed through annotation. Active learning selects uncertain examples using an existing model, while core-set and batch methods seek geometric diversity or representativeness (Sener and Savarese, 2017; Ash et al., 2019). Sign-language studies have explored active selection for isolated-sign acquisition and fingerspelling (Bicego et al., 2023; Wang and Nalisnick, 2023), but selection from continuous unlabeled video before a reliable recognizer or gloss inventory exists remains underexplored.

Sign languages suggest a natural label-free alternative. Hand configuration, location, movement, coordination, and non manual activity can be approximated from pose without manual labels (Börstell, 2023; Kezar et al., 2023). One might therefore spend the annotation budget on clips that add underrepresented articulatory patterns. The intuition is appealing, but coverage need not translate into annotation utility, measured here as downstream task performance under a fixed budget: learning may benefit more from repeated observations of common, representative structures than from maximizing novelty.

We test this hypothesis on PHOENIX14T (Camgoz et al., 2018) and Isharah (Alyami et al., 2026), hiding labels during selection and revealing them only for purchased clips. Pose-derived articulatory coverage, as operationalized here, is generally weak across both sign languages and both tasks, rarely outperforming random sampling. We then ask why. Coverage systematically favors locally atypical clips, yet removing the sparsest candidates recovers only part of the performance gap. A privileged gloss-diversity selector provides the com-

plementary case: it performs poorly even while selecting comparatively typical clips. The evidence therefore points to a broader mismatch between the coverage objectives evaluated here and downstream annotation utility, rather than a failure of pose representation alone.

2 Controlled Study

2.1 Retrospective annotation protocol

Let the unlabeled pool be $\mathcal{U} = \{u_1, \dots, u_N\}$, with estimated annotation cost c_i for clip u_i . Each selector chooses a subset whose total cost does not exceed budget B . Labels remain inaccessible until selection is complete. PHOENIX14T contains 7,096 training clips from studio weather broadcasts; Isharah contributes a 10,000-clip signer-independent pool recorded with smartphones in varied settings. All budgets use a common $6\times$ duration-based annotation-time multiplier, which holds the cost convention fixed across tasks (Appendix G gives the exact hour-to-percent mapping). Recognition budgets span 5 to 20 percent of pool cost: 2.8/5.5/11.0 annotation hours for PHOENIX14T and 7.1/14.2/28.4 for Isharah. Translation follows the same schedule for Isharah and a finer 0.5/1/2/5-hour sweep for PHOENIX14T. Clip cost is proportional to duration, and each budget is defined as a fixed fraction of the total pool cost. Changing the recognition rate from $6\times$ to $40\times$ therefore rescales the reported annotation hours but does not change which clips are affordable, their greedy gain-per-hour ordering, or the subsets ultimately selected. The experimental results are therefore unchanged under this alternative rate (Appendix G). Because cost is duration-based, selectors that prefer shorter clips purchase more clips at the same budget (Table 10). These budgets target severe low-resource conditions in which selection strategy can substantially affect the composition of the labeled subset.

2.2 Selection strategies

Baselines and geometric selection. Random sampling selects clips until the budget is exhausted. Source-stratified random allocates budget across recording sources first, then samples within each; it is a source-imbalance control rather than a primary strategy, so it appears in Appendix A rather than Table 1. For geometric selection, each normalized pose sequence is projected to a generic embedding z_i ; greedy k -center spans the embedding space

(Sener and Savarese, 2017), while facility location favors clips similar to many others, emphasizing dense, representative regions.

Articulatory coverage. Our selection objective favors clips that add signing patterns not yet well represented in the labeled set. Because labels are unavailable during selection, we approximate six articulatory channels from pose: hand configuration, orientation, location, movement, two-hand coordination, and non-manual activity. Each channel is clustered into a separate discrete codebook, so each clip can be represented by the codewords assigned to its frames. Selection can then favor clips that introduce articulatory patterns that are missing or underrepresented in the current labeled set.

The objective incorporates three considerations. First, repeated observations of the same codeword should have diminishing value, so additional coverage contributes progressively less. Second, rare codewords should receive greater weight because common patterns are easier to acquire. Third, clips with poor pose tracking should contribute less because their assigned codewords are more likely to reflect measurement error. We capture these effects with a logarithmic coverage term, a codeword rarity weight r_{gk} , and a clip-level pose-quality score q_i :

$$F_{\text{art}}(\mathcal{S}) = \sum_g \frac{\omega_g}{K_g} \sum_{k=1}^{K_g} r_{gk} \ell_{gk}(\mathcal{S}),$$

$$\ell_{gk}(\mathcal{S}) = \log \left(1 + \sum_{u_i \in \mathcal{S}} q_i n_{igk} \right),$$

where K_g is the number of codewords in channel g , ω_g is its channel weight, and n_{igk} is the capped number of frames in clip i assigned to codeword k . The logarithm gives diminishing returns and makes F_{art} submodular, which supports greedy optimization. The rarity weight r_{gk} increases as a codeword becomes less frequent in the pool, while q_i is floored at 0.1 so that poorly tracked clips are down-weighted rather than excluded. Closed-form definitions of both weights are provided in Appendix G.

We evaluate three nested selectors based on this objective: *articulatory only*, which uses F_{art} alone; *articulatory + repr.*, which adds the facility-location term F_{rep} ; and the *full objective*, $0.60F_{\text{art}} + 0.30F_{\text{rep}} + 0.10F_{\text{src}}$, which also rewards coverage across recording sources. The weights are fixed a priori and are not tuned on any evaluation set. All three selectors use greedy selection by marginal gain per purchased hour.

Model-dependent and privileged comparisons.

To separate a weakness of the coverage objective from one of the pose representation, we also compare against selectors with stronger signals. Entropy sampling and BADGE (Ash et al., 2019) score the pool with a recognizer trained on a random seed set costing 2% of the target annotation budget, making them feasible but not label-free. Privileged diagnostics read labels unavailable at selection time: text facility location operates on hidden translations, and the gloss oracle favors gloss types not yet covered, testing whether coverage helps when defined over linguistic content rather than pose-derived codewords.

Density diagnostic. We measure each clip’s mean cosine distance to its 20 nearest neighbors in the generic pose embedding; larger values indicate sparser local regions. To control for duration, we compare selector density within duration quartiles. We then rerun articulatory coverage after excluding the sparsest 10% or 20% of candidates, leaving the objective, greedy procedure, and budget unchanged.

2.3 Tasks and evaluation

Gloss recognition uses a lightweight pose model with a 256-dimensional projection, temporal convolution, Transformer encoding, and CTC output. Translation uses a matched pose encoder with an autoregressive text decoder. Within each setting, all selectors share the same architecture, optimization, and decoding settings; Appendix G gives full implementation details. The recognition vocabulary contains only glosses present in the purchased subset: the CTC output layer can only emit trained-on glosses, so a reference gloss absent from the purchased subset is scored as an error at evaluation time, a deliberate measure of one consequence of the annotation-selection decision. We report word error rate (WER; lower is better) for recognition and BLEU (higher is better) for translation, with mean and standard deviation over three seeds; privileged diagnostics use one run. The study concerns selector ranking under severe label scarcity rather than state-of-the-art absolute accuracy.

3 Results

3.1 Articulatory coverage is generally weak

Table 1 summarizes the largest-budget results and Figure 1 shows every recognition budget as a difference from random; Appendix A gives the under-

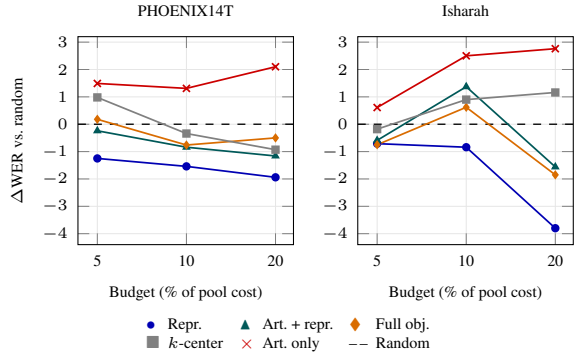


Figure 1: Recognition WER relative to random sampling at each budget, where negative values indicate lower WER than random and the dashed line denotes the random baseline. Articulatory-only selection has higher mean WER than random at every budget on both corpora, while adding representativeness lowers mean WER in each case. Representativeness alone has lower mean WER than random at every budget. Values are computed from Appendix Tables 3 and 4. The privileged gloss oracle is omitted because it falls outside the plotted range, reaching up to +9.7 WER relative to random on Isharah; see Table 1.

lying absolute numbers. On PHOENIX14T recognition, WER rises from 77.46 with representativeness to 78.24 with articulatory + representativeness, 78.90 with the full objective, and 81.50 with articulatory-only selection, the same ordering as at 2.8 and 5.5 hours. PHOENIX14T translation is similarly unfavorable: at 5 hours, representativeness reaches 5.68 BLEU, compared with 4.69 for the full objective and 3.84 for articulatory-only selection. Across the six recognition corpus–budget settings (three PHOENIX14T budgets, three Isharah budgets), articulatory-only selection never outperforms random sampling (Appendix A); on translation it wins only once, at the largest Isharah budget.

The pattern transfers to Isharah recognition. At 28.4 hours, representativeness obtains 76.24 WER, versus 78.19 for the full objective and 82.80 for articulatory only. At the smallest budget, the strongest methods are effectively tied, but the separation grows as more data are purchased. Active learning (entropy sampling, BADGE) and core-set selection (k -center) are also inconsistent across budgets, whereas facility-location representativeness is more reliable: across the 18 corpus–budget–seed comparisons on recognition, computed from individual per-seed WER values (Appendix Table 5), it has lower WER than random in 16 cases against articulatory-only’s 2. Because runs are matched by seed, we compare the 18 paired WER

Strategy	PHX CSLR WER ↓ (11.0 h)	Isharah CSLR WER ↓ (28.4 h)	PHX SLT BLEU ↑ (5 h)	Isharah SLT BLEU ↑ (28.4 h)
Representativeness	77.46 ± 0.2	76.24 ± 2.0	5.68 ± 0.2	6.75 ± 0.2 [†]
<i>k</i> -center	78.47 ± 0.8	81.20 ± 2.2	2.84 ± 0.1	6.93 ± 0.8
Articulatory + repr.	78.24 ± 0.8 [†]	78.48 ± 1.2 [†]	4.62 ± 0.3	5.66 ± 0.3
Full obj. (art+repr+src)	78.90 ± 0.6	78.19 ± 1.7 [†]	4.69 ± 0.5	6.90 ± 0.4 [†]
Random	79.40 ± 1.5	80.04 ± 3.0 [†]	4.40 ± 0.5	5.40 ± 0.9 [†]
Articulatory only	81.50 ± 0.3	82.80 ± 0.6	3.84 ± 0.6	5.88 ± 0.4 [†]
Text facility location (priv.)	79.19	75.13	3.81	3.79
Gloss coverage oracle (priv.)	85.33	89.77	3.89	2.88

Table 1: Results at the largest budget in each setting. PHX denotes PHOENIX14T. The three articulatory rows are nested: articulatory coverage alone, then + representativeness, then + source coverage, corresponding to the full objective with weights 0.60/0.30/0.10. Bold indicates the numerically best feasible selector by mean and does not imply statistical significance. [†] indicates a mean whose ± 1 standard-deviation interval over three seeds overlaps that of the best method; unmarked rows do not overlap the best under this criterion. For Isharah SLT, every method except articulatory + repr., including random, falls within this range of the best, so the leading methods cannot be reliably ranked. Italic indicates an oracle diagnostic that requires labels and is unavailable during unlabeled selection. Entropy sampling and BADGE are feasible baselines reported in Appendix A.

differences directly. Representativeness outperforms random under both tests, with a two-sided sign-test $p = 0.0013$ and Wilcoxon signed-rank $p < 10^{-4}$, corresponding to a mean WER reduction of 1.68 points. Articulatory-only shows the opposite pattern, with higher WER than random under both the sign test ($p = 0.0013$) and Wilcoxon signed-rank test ($p = 0.00025$), with a mean increase of 1.80 points. These 18 pairs are not fully independent because they reuse three seeds across nested budgets. We therefore use the paired tests to summarize the consistency of the observed win counts, rather than as evidence from 18 independent experimental replicates.

3.2 Coverage is not annotation utility

A weak pose representation could explain the pattern above on its own, so the next diagnostic removes pose from the selector entirely and substitutes exact lexical identity. At the 11-hour PHOENIX14T budget the privileged gloss oracle acquires 1,085 of 1,089 gloss types yet yields 85.33 WER, compared with 79.40 for random and 77.46 for representativeness. This shows that lexical type coverage, like articulatory coverage, need not correspond to annotation utility; the failure is not specific to pose descriptors. On Isharah, oracle WER is 89.77, 9.73 points worse than random. Even with exact lexical categories, maximizing type breadth can sacrifice the frequency, context, and repetition a type needs to be learned from.

The selected subsets show the same disconnect (Appendix Table 10, which also reports clip counts). At the 10% PHOENIX14T budget, articulatory cov-

erage acquires more gloss types than representativeness but produces higher WER, while representativeness purchases more mutually similar clips. Breadth and redundancy therefore need not track annotation utility.

3.3 Representativeness is strongest in three of four settings

Facility-location representativeness has the best mean result for recognition on both corpora and for PHOENIX14T translation. Isharah translation is less conclusive: at the largest budget, *k*-center, the full objective, and representativeness obtain 6.93, 6.90, and 6.75 BLEU, respectively, while random obtains 5.40 ± 0.9 . Under the descriptive variability criterion in Table 1, all of these methods overlap the best method’s range, whereas articulatory + repr. does not. The smaller-budget BLEU values are also low relative to their run-to-run variation. We therefore do not assign a clear ranking among the leading methods for Isharah translation.

4 Why Does Coverage Fail?

Coverage favors locally atypical clips.

Articulatory-only selection (Table 2) is sparser than the pool in every duration quartile of both corpora, most clearly on Isharah, while random stays near baseline: coverage systematically favors unusual pose configurations rather than inheriting a duration bias.

Filtering helps but does not rescue coverage. If atypicality were the sole reason coverage fails, removing the sparsest candidates should eliminate

Corpus	Duration-controlled Δ kNN distance		Articulatory-only WER at largest budget		
	Art. coverage	Random	Unfiltered	Best filtered	Random
PHOENIX14T	[+0.017, +0.034]	[−0.009, +0.007]	81.50	79.94	79.40
Isharah	[+0.058, +0.092]	[−0.006, +0.014]	82.80	81.99	80.04

Table 2: Density and filtering diagnostics. Δ kNN distance is given as the range over the four duration quartiles, each measured against its own quartile pool mean; positive values indicate locally sparser clips. Articulatory coverage is sparser than the pool in every quartile of both corpora, whereas random straddles zero. The WER columns report articulatory-only selection before filtering, after removing the sparsest 10–20% of candidates, and random for reference.

the deficit. Instead it produces only partial recovery: articulatory-only WER improves from 81.50 to 79.94 on PHOENIX14T and 82.80 to 81.99 on Isharah, yet the best filtered variant remains at or above random’s mean WER in all six recognition settings (Appendix Table 9). Atypical-example selection therefore contributes to the weakness of coverage but does not fully explain it.

Diversity can fail even without atypicality. The privileged gloss oracle is a complementary counterexample: on PHOENIX14T it is the weakest recognition selector despite being denser than the pool in three of four quartiles (Appendix Table 8) and acquiring nearly the complete gloss inventory. Even direct access to lexical categories can sacrifice the repetitions or combinations needed for learning.

Supporting diagnostics. Two additional diagnostics support this interpretation (Appendix D): random selection saturates articulatory codebooks quickly, and coverage-based selection is more sensitive to pose corruption than representativeness is. The learned channels also show low signer leakage and strong temporal coherence (Appendix E), arguing against the explanation that the representation is dominated by signer identity or frame-level noise. Together, these diagnostics point to a mismatch between the coverage objectives evaluated here and annotation utility, rather than a single implementation failure.

5 Discussion

The results highlight a distinction between coverage and learnability. A selector can acquire many rare articulatory or lexical types without purchasing enough repeated, representative evidence for those types to be learned effectively. This helps explain why filtering atypical clips improves articulatory coverage only partially, and why even near-complete lexical coverage can perform poorly. Facility-location representativeness provides the

strongest results in three of the four corpus–task settings, while Isharah translation does not yield a clear ranking among the leading methods.

These results are specific to the low-resource, pose-based setting evaluated here. Both downstream models are small (4.1M parameters), so the observed ranking reflects the value of different selection strategies under a fixed annotation budget for models of this scale. Larger models, particularly those operating on RGB input, may be able to exploit forms of diversity that these models cannot, and a coverage-selected subset could therefore rank differently in that setting. We therefore interpret our results as evidence that coverage is not a reliable default for low-resource, pose-based annotation selection, rather than as a claim that coverage can never be beneficial.

6 Conclusion

Selecting clips to maximize pose-derived coverage or lexical type coverage, as operationalized by the objectives tested here, is not reliably better than random sampling in low-resource sign-language annotation. Across two corpora and two tasks, pose-based coverage underperforms, and privileged lexical coverage does not rescue the underlying idea. This does not imply that all diversity objectives fail; our conclusion is limited to the coverage objectives evaluated here. Selection policies should therefore be justified by annotation utility, with representativeness treated as a strong empirical baseline rather than coverage as a default objective.

Limitations

The study uses sentence-segmented benchmarks rather than unsegmented community corpora, and fixed budget units rather than a prospective annotator study. Segmentation is also part of the data collection process, not just a fixed property of the benchmark. Our selectors operate on clips

with predefined boundaries, whereas a corpus built from continuous recordings must first decide where those boundaries fall. Coverage-based selection may behave differently when segmentation itself is part of the selection process. Both downstream models are small and pose-based, so conclusions concern selector ranking under severe label scarcity and may not transfer to large RGB models (see Section 5). PHOENIX14T is a narrow weather domain, Isharah provides only one additional sign language, and the gloss oracle tests type coverage rather than all diversity objectives.

The pose pipeline also exposes a measurement-quality concern: dominant-hand detection varies substantially across signers, with a 0.365 spread on Isharah, so a noise-sensitive selector may allocate annotation partly by measurement quality rather than signing content. We do not infer a causal effect on translation; the signer-level correlation with BLEU is not statistically significant ($\rho = 0.42$, $p = 0.27$, $n = 9$).

All seed-level results use three seeds. This is too few for reliable statistical testing of individual cells in Table 1, so small differences are described as within run-to-run variation rather than assigned per-cell significance tests. Section 3.1 instead uses the seed-matched design to analyze the 18 paired differences across conditions. These paired comparisons provide more information than any single three-seed cell, but they are still not fully independent because the same seeds are reused across nested budgets. We therefore interpret the resulting tests as evidence about the consistency of the paired differences, not as inference from 18 independent replicates. The $6\times$ multiplier used to price the annotation budgets (Appendix G) is an assumption rather than a measured glossing rate. Although the multiplier does not affect clip selection or any within-task comparison, it does affect how the reported hours should be interpreted across tasks. If glossing requires approximately $40\times$ real time while translation requires $6\times$, then recognition and translation experiments reported at the same nominal number of hours do not correspond to the same annotator effort. The hour budgets should therefore be interpreted within each task rather than as directly comparable annotation costs across tasks.

Acknowledgments

We thank anonymous reviewers for their valuable feedback.

References

- Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaiifeer, Yazeed Alharbi, and Yaser Alon-aizan. 2026. Isharah: A large-scale multi-scene dataset for continuous sign language recognition. *IEEE Transactions on Multimedia*.
- Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.
- Manuele Bicego, Manuel Vázquez-Enríquez, and José L Alba-Castro. 2023. Active class selection for dataset acquisition in sign language recognition. In *International Conference on Image Analysis and Processing*, pages 304–315. Springer.
- Carl Börstell. 2023. Extracting sign language articulation from videos with mediapipe. In *Proceedings of the 24th nordic conference on computational linguistics (nodalida)*, pages 169–178.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793. Ieee.
- Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4all open research sign language translation datasets. In *2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021)*, pages 1–5. IEEE.
- Mirella De Sisto, Vincent Vandeghinste, Santiago Egea-Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487.
- Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743. IEEE.
- Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023. Improving sign recognition with phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Shuai Wang and Eric Nalisnick. 2023. Active learning for multilingual fingerspelling corpora. *arXiv preprint arXiv:2309.12443*.

A Complete Results at All Budgets

Table 1 in the main text reports the largest budget in each setting. Tables 3 to 7 report results at every budget.

Strategy	2.8 h	5.5 h	11.0 h
Representativeness	88.73 ± 0.5	85.05 ± 0.3	77.46 ± 0.2
Articulatory + repr.	89.74 ± 0.6	85.75 ± 0.4	78.24 ± 0.8
k -center	90.96 ± 0.5	86.25 ± 0.6	78.47 ± 0.8
Full without quality wt.	89.68 ± 1.2	85.54 ± 0.1	78.42 ± 0.4
Full obj. (art+repr+src)	90.16 ± 0.5	85.83 ± 0.2	78.90 ± 0.6
Full without rare wt.	90.43 ± 0.5	86.10 ± 0.6	79.08 ± 0.4
Random	89.98 ± 1.2	86.59 ± 0.8	79.40 ± 1.5
BADGE	89.91 ± 1.7	85.87 ± 0.7	79.71 ± 0.8
Source-stratified random	89.93 ± 0.6	87.66 ± 1.0	81.13 ± 1.1
Entropy sampling	—	87.10 ± 0.8	81.23 ± 0.8
Articulatory only	91.47 ± 0.8	87.90 ± 1.1	81.50 ± 0.3
<i>Text facility loc.</i>	88.50	85.11	79.19
<i>Gloss oracle</i>	92.98	89.57	85.33

Table 3: PHOENIX14T gloss recognition, WER $\downarrow$. Entropy sampling is defined at the 10 and 20 percent budgets only.

Strategy	7.1 h	14.2 h	28.4 h
Representativeness	90.78 ± 0.5	88.15 ± 0.3	76.24 ± 2.0
Full obj. (art+repr+src)	90.74 ± 0.6	89.61 ± 0.3	78.19 ± 1.7
Articulatory + repr.	90.90 ± 0.3	90.36 ± 0.7	78.48 ± 1.2
BADGE	91.03 ± 0.4	89.84 ± 1.5	79.45 ± 1.1
Source-stratified random	91.61 ± 1.5	89.63 ± 0.3	79.70 ± 4.1
Random	91.49 ± 0.4	88.99 ± 0.6	80.04 ± 3.0
k -center	91.31 ± 0.9	89.89 ± 0.4	81.20 ± 2.2
Entropy sampling	—	89.82 ± 2.3	82.40 ± 4.3
Articulatory only	92.10 ± 0.9	91.49 ± 1.0	82.80 ± 0.6
<i>Text facility loc.</i>	90.37	89.06	75.13
<i>Gloss oracle</i>	91.56	90.98	89.77

Table 4: Isharah gloss recognition, WER $\downarrow$. At 7.1 h the strongest three strategies lie within 0.2 WER and should be read as tied.

B Quartile-Level Density by Selector

Section 4 states that the gloss oracle is denser than the duration-matched pool in three of four quartiles on PHOENIX14T. Table 8 gives the full per-quartile breakdown behind that claim and behind Table 2, for every selector discussed in the main text, not only articulatory-only and random.

Budget	Seed 0	Seed 1	Seed 2
<i>PHOENIX14T, representativeness — random (WER)</i>			
2.8 h	−0.29	−1.65	−1.79
5.5 h	−0.51	−2.64	−1.47
11.0 h	+0.05	−2.83	−3.04
<i>Isharah, representativeness — random (WER)</i>			
7.1 h	−0.81	−0.54	−0.79
14.2 h	−1.19	+0.21	−1.52
28.4 h	−2.46	−4.95	−3.97
<i>PHOENIX14T, articulatory-only — random (WER)</i>			
2.8 h	+2.16	+1.41	+0.91
5.5 h	+0.93	+1.52	+1.49
11.0 h	+3.58	+1.55	+1.17
<i>Isharah, articulatory-only — random (WER)</i>			
7.1 h	+1.82	−0.13	+0.13
14.2 h	+2.07	+4.28	+1.17
28.4 h	+5.28	−1.36	+4.36

Table 5: The 18 corpus–budget–seed comparisons behind the win counts in Section 3.1, as (selector WER — random WER); negative means the selector beats random at that seed. From the per-seed runs underlying Tables 3 and 4.

Strategy	0.5 h	1 h	2 h	5 h
Representativeness	4.01 ± 0.1	4.08 ± 0.2	4.49 ± 0.2	5.68 ± 0.2
Random	2.59 ± 0.3	3.87 ± 0.5	3.57 ± 0.5	4.40 ± 0.5
Articulatory + repr.	1.99 ± 0.2	3.57 ± 0.1	3.84 ± 0.4	4.62 ± 0.3
Full obj. (art+repr+src)	1.60 ± 0.2	2.62 ± 0.1	3.06 ± 0.2	4.69 ± 0.5
Articulatory only	0.93 ± 0.1	1.12 ± 0.2	1.91 ± 0.2	3.84 ± 0.6
k -center	0.87 ± 0.2	1.38 ± 0.1	1.85 ± 0.2	2.84 ± 0.1

Table 6: PHOENIX14T translation, BLEU $\uparrow$.

C What Each Strategy Purchases

Table 10 describes the selected subsets at the 10-percent PHOENIX14T budget rather than the models trained on them.

Articulatory coverage acquires *more* gloss types than representativeness, 497 against 451, and still produces higher error. At this 10-percent (5.5-hour) budget the gloss oracle acquires 1,024 of 1,089 types, fewer than at the 20-percent (11-hour) budget reported in Section 3.2 (1,085 of 1,089), and is the weakest method overall at both. Meanwhile the strongest selector is one of the most redundant methods, 0.527 against 0.448 for random: it deliberately purchases clips that resemble one another. Type breadth and redundancy both point the opposite way from annotation utility.

Representativeness acquires fewer distinct types than articulatory-only selection, but provides more repeated evidence for each one: 13.37 occurrences per type on average, compared with 11.33. In other words, representativeness favors depth, whereas articulatory coverage favors breadth. The gloss or-

Strategy	7.1 h	14.2 h	28.4 h
<i>k</i> -center	0.92 $\pm$ 0.2	1.87 $\pm$ 0.4	6.93 $\pm$ 0.8
Full obj. (art+repr+src)	1.19 $\pm$ 0.3	2.66 $\pm$ 0.5	6.90 $\pm$ 0.4
Representativeness	1.17 $\pm$ 0.4	1.92 $\pm$ 0.5	6.75 $\pm$ 0.2
Articulatory only	1.24 $\pm$ 0.5	2.35 $\pm$ 0.5	5.88 $\pm$ 0.4
Articulatory + repr.	1.12 $\pm$ 0.2	2.11 $\pm$ 0.3	5.66 $\pm$ 0.3
Random	1.24 $\pm$ 0.2	2.56 $\pm$ 0.5	5.40 $\pm$ 0.9

Table 7: Isharah translation, BLEU $\uparrow$. Discussed in Section 3.3; scores below 14.2 h have standard deviations too large relative to the spread to interpret the ordering.

Strategy	Q1	Q2	Q3	Q4
<i>PHOENIX14T</i>				
Articulatory only	+0.017	+0.034	+0.025	+0.022
Full obj. (art+repr+src)	+0.006	+0.022	-0.002	-0.007
Representativeness	+0.012	-0.003	-0.104	—
<i>k</i> -center	+0.012	+0.116	—	—
Gloss oracle (priv.)	-0.007	+0.003	-0.017	-0.021
Random	+0.004	-0.007	+0.007	-0.009
<i>Isharah</i>				
Articulatory only	+0.058	+0.058	+0.061	+0.092
Full obj. (art+repr+src)	+0.035	+0.046	+0.060	+0.041
Representativeness	+0.049	+0.025	-0.015	-0.075
<i>k</i> -center	+0.044	+0.167	—	—
Gloss oracle (priv.)	+0.035	+0.035	+0.042	+0.076
Random	+0.014	+0.009	-0.006	-0.002

Table 8: Mean kNN distance in the generic pose embedding, expressed as a difference from each quartile’s own pool mean, for every selector’s purchases at the 10-percent budget. Positive values indicate clips that are sparser than typical clips of the same duration. — marks a quartile with fewer than 10 purchased clips.

acle shows the same trade-off at the extreme. It covers 1,024 of 1,089 types, yet provides only 6.06 tokens per type, roughly half the support of any feasible selector, and is also the weakest method for recognition despite its near-complete type coverage. The same pattern appears at the 20-percent budget, where representativeness provides 20.16 tokens per type compared with 17.38 for articulatory-only selection. These results suggest that coverage alone is insufficient: recognition also benefits from repeated evidence for the types that are selected.

D Supporting Diagnostics

We use two additional diagnostics to examine the behavior of articulatory coverage. The saturation experiment varies the hand-configuration codebook size as a diagnostic sweep. The main hand-configuration channel used everywhere else in the paper, including every row of Table 1, has 64 code-words; the 40, 150 and 400 settings below are diagnostic sweeps only and are not used by the main selector. The corruption experiment measures how stable each selected subset remains when detected

Budget	Unfiltered	Dens10	Dens20	Random
<i>PHOENIX14T</i>				
2.8 h	91.47	91.81	91.94	89.98
5.5 h	87.90	87.73	86.92	86.59
11.0 h	81.50	79.94	80.19	79.40
<i>Isharah</i>				
7.1 h	92.10	93.20	91.61	91.49
14.2 h	91.49	90.21	90.05	88.99
28.4 h	82.80	81.99	82.02	80.04

Table 9: Articulatory-only WER (mean over three seeds) before filtering, after removing the sparsest 10% and 20% of candidates, and for random selection, at all six recognition budgets. The best filtered variant remains at or above random’s mean WER in every setting.

hand landmarks are randomly dropped.

E Validating the Articulatory Channels

Any codebook fitted to pose produces clusters. Whether those clusters separate articulation rather than the people producing it is a separate question, and the failure mode is concrete: clustering handshape descriptors can converge on hand size, camera distance or seating position. Such clusters would still yield a coverage curve, so the problem is invisible downstream.

We measure the excess normalized mutual information between cluster assignment and signer identity, relative to a permutation null. Raw NMI is biased upward when clusters are many and signers few, so the null distribution is the only meaningful reference. We also test whether cluster identity persists across adjacent frames, as articulatory state should, and whether the handshape and location channels are redundant.

The thresholds are calibrated on synthetic pose where the answer is known by construction: one set in which handshape is drawn independently of signer, and one in which handshape is a deterministic function of signer. Without this calibration a passing score cannot be interpreted. The 0.05 signer-leakage threshold sits roughly an order of magnitude above the independent-control value (0.007) and well below the signer-determined one (0.698), so both corpora’s observed values (0.018, 0.037) are unambiguously on the independent side.

F Pose Reliability Across Signers

The measurement-disparity observation in the Limitations section rests on the measurements below. The Isharah pose set measured here spans 18 signers with 500 to 2,000 clips each, so its

Strategy	Clips	Types	Gloss cov.	H_{src}	Redund.	Tokens/type	Types ≥ 5
Random	702	476	0.439	0.846	0.448	11.73	181
Source-strat.	705	486	0.448	0.986	0.460	11.35	189
k -center	1491	440	0.406	0.831	0.530	13.90	178
Representativeness	1427	451	0.416	0.835	0.527	13.37	188
Articulatory only	827	497	0.458	0.837	0.453	11.33	208
Full obj. (art+repr+src)	1224	491	0.453	0.932	0.517	11.98	192
<i>Text facility loc.</i>	1280	505	0.465	0.835	0.524	11.74	186
<i>Gloss oracle</i>	773	1024	0.944	0.853	0.485	6.06	276

Table 10: Properties of the purchased subset at the 10 percent PHOENIX14T budget. H_{src} is normalized entropy over recording sources; redundancy is mean nearest-neighbor cosine similarity within the selected set. Tokens/type and Types ≥ 5 measure repetition within the purchase itself, not coverage of the pool. Gloss coverage is measured after selection and was not available to any label-free selector. Entropy sampling and BADGE are omitted: their purchases depend on a seed model that was not retained, so they cannot be reproduced exactly. Articulatory only and Full obj. were re-run for this table; both shifted by $\leq 2\%$ from the originally reported run.

Budget	40 codewords		150 codewords		400 codewords	
	Random	Gain	Random	Gain	Random	Gain
0.5 h	1.000	0.000	0.997	0.003	0.847	0.143
1 h	1.000	0.000	1.000	0.000	0.942	0.055
2 h	1.000	0.000	1.000	0.000	0.981	0.019
5 h	1.000	0.000	1.000	0.000	0.997	0.003
10 h	1.000	0.000	1.000	0.000	0.999	0.001
40 h	1.000	0.000	1.000	0.000	1.000	0.000

Pose-corruption stability: duration-weighted overlap with clean selection

Strategy	10% dropout	20% dropout
Articulatory only	0.622	0.559
Full obj. (art+repr+src)	0.746	0.693
Representativeness	0.822	0.777

Table 11: Supporting diagnostics for articulatory coverage. The upper panel reports the fraction of the hand-configuration inventory covered by uniform random selection and the additional coverage obtained by greedy coverage selection. The 40-, 150-, and 400-codeword settings are diagnostic sweeps; the main selector uses 64 codewords. The lower panel reports duration-weighted overlap between selections made from clean and corrupted pose, where higher values indicate greater stability.

Test	PHOENIX14T	Isharah	Threshold
Signer leakage (excess NMI)	0.018	0.037	< 0.05
Temporal coherence (lift)	$8.97\times$	$10.46\times$	$> 3\times$
Channel independence (NMI)	0.047	0.048	< 0.30
<i>Synthetic calibration, excess NMI</i>			
handshape $\perp$ signer		0.007	
handshape = signer		0.698	

Table 12: Validation of the articulatory channels. Both corpora sit near the independent control and far from the signer-determined one. This establishes that the channels are not degenerate; it does not establish that they correspond to linguistic categories.

spread is not a small-sample artifact; the signer-independent training pool draws on 13 of them. On PHOENIX14T the extremes are the two smallest signers; restricted to signers with more than 500 clips the spread is 0.107.

Correlating per-signer detection rate with per-

Per-frame detection rate	PHOENIX14T	Isharah
	<i>studio</i>	<i>smartphone</i>
Dominant hand, mean	0.771	0.836
Non-dominant hand, mean	0.720	0.712
Face	0.876	0.963
Dominant hand, worst signer	0.578	0.608
Dominant hand, best signer	0.834	0.974
Spread across signers	0.256	0.365

Table 13: Landmark detection rates. Upsampling PHOENIX14T frames by $2\times$, $3\times$ and $4\times$ leaves detection essentially unchanged (0.691 to 0.687), so the limit is detail absent from the source rather than resolution. Smartphone footage recorded in unconstrained environments detects better than studio broadcast.

signer BLEU across all runs gives Spearman $\rho = 0.42$, $p = 0.27$ over nine signers. We report the input disparity, which is directly measured, and do not claim it causes worse translation.

G Implementation Details

Pose and normalization. Landmarks are extracted with MediaPipe Holistic (Lugaresi et al., 2019). Each frame is translated to the shoulder midpoint, scaled by shoulder width, and rotated so the shoulder axis is horizontal, so that body roll does not produce different location codewords for the same articulation. Coordinates are clipped to $[-5, 5]$ and median filtered with window five; gaps of at most five frames are linearly interpolated and longer gaps remain masked.

Articulatory channels. Six groups with codebook sizes 64 (hand configuration), 16 (orientation), 24 (location), 32 (motion), 8 (coordination) and 16 (nonmanual). Each hand is translated to its wrist, scaled by palm size, rotation aligned, and left hands are mirrored into the right-hand coordinate frame so the same configuration on either hand maps to one codeword. For handshake, orientation and location only frames below the clip’s 60th speed percentile are used, since those parameters describe held configurations rather than transitions between them. Codebooks are fitted with MiniBatch k -means on descriptors standardized by median and interquartile range, sampling at most 200 descriptors per clip and 200,000 across the pool. Frames in which a hand is not detected are excluded before quantization; including them creates a codeword meaning “hand not found”, which a coverage objective then treats as a rare pattern worth purchasing.

Pose quality is $q_i = \text{clip}(\frac{v_i + a_i + r_i}{3}, 0.1, 1.0)$, where v_i is the fraction of frames with a detected hand, a_i is mean landmark availability across all tracked streams, and r_i is temporal continuity (the fraction of adjacent frames whose hand-validity does not flip); the floor at 0.1 down-weights poorly tracked clips without excluding them. The rarity weight is $r_{gk} = \min(5, \sqrt{\frac{\text{med}_g}{f_{gk}}})$, where f_{gk} is the pool frequency of codeword k in channel g and med_g is the median codeword frequency within that channel.

Downstream models. Gloss recognition uses a 4.1M-parameter model: frame projection to 256 dimensions, two temporal convolution blocks (kernel 5, stride 2), four pre-norm Transformer encoder layers with four heads and feed-forward width 1024, and a CTC output. Sequences longer than 320 frames are resampled. Translation uses a matched encoder with an autoregressive text decoder. Training uses AdamW at 3×10^{-4} , weight decay 10^{-2} , 5 percent warmup, cosine decay, at most 60 epochs

with early stopping at patience 8, and an effective batch of 32.

Cost model. The pipeline supports two annotation-time multipliers, $6\times$ for translation and $40\times$ for gloss, the latter anchored on the How2Sign corpus’s reported figure of approximately one annotator-hour per 90 seconds of video for gloss annotation (Duarte et al., 2021). Every budget reported in this paper, for both recognition and translation, uses the $6\times$ translation multiplier. Recognition budgets (2.8/5.5/11.0 h for PHOENIX14T, 7.1/14.2/28.4 h for Isharah) are 5/10/20 percent of each corpus’s pool cost under that rate; Isharah translation reuses the same schedule. PHOENIX14T translation instead uses a finer 0.5/1/2/5-hour sweep, 0.9/1.8/3.6/9.1 percent of its pool cost. We use a common duration-based cost axis rather than task-specific annotation rates so that selector comparisons are not confounded by a change of cost model across tasks; this is a controlled-study convention, not a claim about the true relative annotation cost of glossing and translation. The $40\times$ gloss rate is not used to price any of the reported experiments. Budgets are expressed as the same fraction of pool annotation cost on both corpora rather than the same absolute hours, since Isharah’s pool is 2.6 times larger.

Why the multiplier cannot affect the reported results. Pricing gloss recognition at the translation rate understates the true annotation cost, but it does not affect the selected subsets or downstream results. Let the multiplier be m , so clip i has cost $c_i = m d_i$, where d_i is its duration, and let the annotation budget be a fixed fraction β of the total pool cost, $B = \beta m \sum_j d_j$. The feasibility constraint $\sum_{i \in \mathcal{S}} m d_i \leq \beta m \sum_j d_j$ then reduces to $\sum_{i \in \mathcal{S}} d_i \leq \beta \sum_j d_j$, so the feasible subsets are independent of m . The greedy selection rule is invariant for the same reason: ranking candidates by marginal gain per purchased hour, $\Delta F/(m d_i)$, gives the same ordering as $\Delta F/d_i$ for any $m > 0$. Entropy sampling and BADGE are also unchanged because their seed sets are priced as a fixed 2 percent of the same target budget. Replacing the $6\times$ rate with the $40\times$ gloss rate therefore leaves every purchased subset and downstream result unchanged; it only rescales the reported annotation hours by $40/6$. The corresponding recognition budgets would be 18.7/36.7/73.3 h for PHOENIX14T and 47.3/94.7/189.3 h for Isharah. This invariance follows directly from the budget definition and selection rule rather than from an empirical sensi-

tivity analysis. The rate assumption still affects comparisons of recognition and translation budgets in absolute annotator hours, since the two tasks should not be interpreted as having the same real annotation cost.

Label isolation. The constraint that selection never sees a label is enforced by a static analysis test that parses each selection module and fails if it reads a gloss or translation field, by attribute or by subscript. Naming an annotation type in the cost model is permitted, since the cost model must know the relative price of a gloss and a translation.