"A Sacred Bird Called the Phoenix". Auditing the most-used Parallel Corpus for German Sign Language Recognition and Translation

Vera Czehmann[1,2], **Shakib Yazdani**[1], **Yasser Hamidullah**[1],
Fabrizio Nunnari[1], **Eleftherios Avramidis**[3,1]
[1]German Research Center for Artificial Intelligence (DFKI GmbH),
[2]Technische Universität Berlin, Berlin, Germany,
[3]alangu GmbH, Cologne, Germany
{vera.czehmann, shakib.yazdani, yasser.hamidullah,
fabrizio.nunnari, eleftherios.avramidis}@dfki.de, avramidis@alangu.de

Abstract

This paper presents an empirical audit of the widely used RWTH-PHOENIX-2014T corpus, examining its suitability as a benchmark for sign language recognition and translation. Through human annotation of the training set and extensive sign-to-text back translation of the test set, we provide detailed statistics that indicate substantial quality issues, including information loss and lexical errors. Automatic scores comparing human sign-to-text back translations to the original speech-transcribed references are remarkably low, suggesting strong translationese effects and substantial paraphrasing, revealing limitations of lexical metrics in adequately scoring translation quality. Replacing the original speech-transcribed references with human sign-to-text back translations while scoring existing sign language translation systems reveals the lack of robustness of system evaluation with lexical metrics against this test set. Our findings highlight risks associated with relying on this corpus for model evaluation and call for more rigorous, linguistically grounded evaluation practices in sign language technology research. The back-translated test set and error annotations are made publicly available.

Keywords: sign language, dataset, corpus analysis

1. Introduction

Sign language technology is in its early stages, and the research community has recently become increasingly active in its efforts to achieve significant progress. Automatic recognition (SLR) and translation of sign language (SLT) are two of the most promising, but also most challenging, application areas. Progress in these areas depends on state-of-the-art machine learning methods, which in turn require sound automatic evaluation protocols and high-quality test sets. Motivated by the belief that evaluation protocols should themselves be critically examined for robustness and validity, we investigate commonly used evaluation practices for research on sign language processing.

We focus on the most widely used test set in German sign language processing research, the RWTH-PHOENIX-2014T corpus (hereafter referred to as "PHOENIX"), which has been used as a benchmark for hundreds of research experiments. Despite its significance to the research community, several limitations of the corpus have already been noted in prior work. We revisit these concerns and provide a more detailed empirical analysis of the corpus, with a particular focus on the test set, its limitations, and the implications for evaluation practice. Beyond the corpus audit itself, we examine benchmark stability under original and sign-to-text back-translated reference conditions. Our goal is to raise awareness of potential weaknesses in current evaluation setups and to encourage more rigorous and linguistically grounded evaluation practices in sign language technology research.

In Section 2 we review relevant background, prior work and resources: we recap standard evaluation protocols, describe the RWTH–PHOENIX–2014/2014T corpus, summarise other published DGS datasets, and outline the main criticisms motivating our audit. The research questions and analysis design are presented in Section 3. The experimental setup, including sampling, human error annotation of the training set, test set back translation,

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 80–92

automatic scoring, and benchmark evaluation is described in Section 4. We then report quantitative findings in Section 5, and conclude with recommendations for evaluation practices and dataset use in Section 6.

2. Background

In this section we summarise common evaluation protocols in sign language translation, describe the RWTH-PHOENIX-2014/2014T corpus and related DGS resources, and review previously noted limitations that motivate our audit.

2.1. Evaluation Protocols

Current SLT development relies heavily on state-of-the-art machine learning methods and models trained on large amounts of parallel data (training sets), i.e. datasets that represent the same content in both source and target language. In machine learning, automatic evaluation protocols are used in order to provide quick feedback during development, and to guide decisions and settings for the creation of the models (Kohavi, 1995; Hastie et al., 2009). Such protocols typically consist of one or more automatic evaluation metrics, and a test set (Landy et al., 1978; Arlot and Celisse, 2010).

The test set is a parallel corpus that is not part of the training data, and therefore also referred to as a "blind" or "hold-out" set. The source side is provided to the trained translation model, and the resulting output is compared against the target side of the test corpus (reference translation), providing a numerical result indicative of the translation performance (Papineni et al., 2002; Rei et al., 2020). These scores allow comparison across models and can inform model selection, or, under comparative ablation analyses, indicate which settings or design decisions should be kept (Och, 2003; Demšar, 2006).

The use of such an evaluation protocol rests on the following assumptions: (a) that the automatic evaluation metric provides a quality judgment that approximates that of users (Callison-Burch et al., 2008; Mathur et al., 2020) and (b) that the test set has the necessary properties so that the findings of the evaluation generalise to the use cases relevant to the audience of the model or the performed analysis (Dwork et al., 2015). In this work, we focus on (b).

An important requirement of a test set is that it is itself of high quality. For machine translation, two properties are particularly important:

- the source content of the test set should be understandable in its own right, for example by a competent human translator, and
- the reference translations should be accurate translations of the source, since automatic metrics rely on them to provide meaningful statistics over model output (Freitag et al., 2020).

These aspects form the basis of our analysis.

2.2. The PHOENIX Corpus

RWTH-PHOENIX-Weather 2014/2014T (Stein et al., 2010; Forster et al., 2012, 2014) is a benchmark corpus for German Sign Language (DGS)-based continuous sign language recognition and translation. It is built from German public TV weather forecasts with simultaneous sign-language interpretation, and is widely used for SLR/SLT research and baselines. The corpus name derives from the name of the TV channel, *Phoenix*, which relays the news of the first channel of the public broadcaster with additional live sign language interpretation. The channel itself has been named after the mythical bird phoenix that rises from its own ashes,[1] symbolising a "media-political correction" and a fresh start for in-depth information.[2]

The signer videos are cropped to the interpreter inset ("interpreter box only") recorded at 25 fps with a frame size of 210×260 pixels, on a controlled grey background with signers wearing dark clothing to ease vision-based processing. The corpus provides sentence- and gloss-level annotations for continuous signing,

[1] Herodotus, Histories 2.73 "*There is also another sacred bird called the phoenix, which I did not myself see except in painting...*"; used for the title of the paper.

[2] History of the broadcaster, Phoenix website, retrieved on 15 Feb 2026.

as well as parallel spoken-language text. On the spoken-language side, the German text is a semi-automatic ASR transcript of the presenter's speech. An open-source speech recognition system was applied to the audio stream of the videos, and the recognition output was then manually corrected by native German speakers. Each SLT instance can therefore be represented as a triplet consisting of video, gloss sequence, and German text.

The latest release (2014T) contains roughly 0.95 million frames, >67k signs from a ∼1k-sign vocabulary, and >2.8k German word types, with standard train/dev/test splits, and the source comprises ∼386 weather editions recorded between 2009–2011. In addition to the corpus itself, the maintainers also distribute precomputed features, scoring tools and documented baselines for both multisigner SLR and SLT setups, which has contributed to making PHOENIX a de-facto standard for quantitative comparisons in sign-language processing.

Citation statistics[3] reveal that the corpus appears to have been used for more than 300 experiments since it was first published in 2010 (Stein et al., 2010), and it is still very actively used in recent publications. Since 2020, when experiments on machine learning over sign languages spiked, more than 30 experiments per year report their results on PHOENIX. Its use is stable over the years, as these numbers are maintained during 2025.

2.3. Other German Sign Language Datasets

To situate PHOENIX within the broader DGS resource landscape, we summarise other published datasets for German Sign Language. The flagship linguistic resource is the **Public DGS Corpus** (MY DGS–annotated; Hanke et al., 2020), released in several waves by the DGS-Korpus project with multi-portal access, rich manual annotations, and even pose tracks. Earlier DGS resources include **SIGNUM**, a signer-independent corpus with 25 native signers, 450 lexemes and 780 sentences (von Agris and Kraiss, 2010), and the **ATIS** Sign Language corpus that provides DGS alongside other languages within an air-travel domain (Bungeroth et al., 2008). Multilingual projects that include DGS comprise **Dicta-Sign**, a four-language corpus (BSL, DGS, GSL, LSF) with harmonised elicitation materials and multi-camera capture (Matthes et al., 2012). Recent additions include **DGS-Fabeln-1**, a parallel collection of German text and DGS fairy-tale interpretations recorded from multiple angles (Nunnari et al., 2024), and larger broadcast-based compilations such as the **TUB Sign Language Corpus Collection** (Avramidis et al., 2025). This overview shows that the prominence of PHOENIX in benchmark-driven SLT research should be understood against the background of other corpora with different properties and intended uses.

2.4. Prior Criticism of PHOENIX

Several limitations of PHOENIX have already been discussed in the literature, including by the creators of the corpus themselves. Stein et al. (2010) criticise the real-time interpreting situation, pointing to bias toward spoken German grammar, omissions in signed output, and limitations for high-quality linguistic analysis and translation evaluation. Forster et al. (2012) note the compactness of the weather domain, motion blur due to fast signing and the 25 fps frame rate, the amount of singleton glosses, and loose spoken–signed parallelism as issues for recognition and translation tasks. Forster et al. (2014) explicitly state that the simultaneous interpreting setup affects spatio-temporal organisation, differs from "real-life" deaf signer language, and that the corpus is not intended primarily for linguistic research. They also mention loose translations and out-of-vocabulary/signature sparsity.

Other work has raised related concerns. Bragg et al. (2019) survey SL datasets and highlight general limitations: small signer pools, constrained domains (such as weather/news), interpreted rather than naturally produced sign language, and mismatch with real-world language use, with PHOENIX as a central example. When reviewing sign language datasets, De Sisto et al. (2022) note that PHOENIX lacks annotation of non-manual features, whereas the fact that the

[3] Google Scholar and Semantic Scholar

annotation guidelines of the existing glosses are not publicly available may raise a transparency issue. Müller et al. (2023) discuss the live interpretation and translationese effects such as omission of information and influence from German grammatical structure, the tokenization and removal of punctuation from the spoken language text, and the glossing quality, stating that "from a scientific point of view achieving higher gloss translation quality on the PHOENIX dataset is near meaningless". Additionally, they note that PHOENIX is overused as a dataset, which is reminiscent of the overuse of MNIST (LeCun et al., 2010) in machine learning, and the WMT14 English-German test set (Bojar et al., 2014) in machine translation of text (Vaswani et al., 2017). A recent study of automatic SLT evaluation (Yazdani et al., 2026) uses PHOENIX to highlight the limitations of text-based metrics.

While not explicitly referring to PHOENIX, the *Sign Language Dataset Compendium* (Kopf et al., 2022; Schulder et al., 2025) excluded this dataset from its selection, possibly because it does not fulfill the curation criterion of containing natural non-interpreted sign language.

Prior SLT experiments run on both PHOENIX and the Public-DGS corpus (e.g. Angelova et al., 2022; Zhu et al., 2023) also indicate a large discrepancy between the automatic evaluation scores and the perceived translation quality, for models trained and tested on the two corpora, with the scores for Public-DGS being much lower than the ones of PHOENIX. This might point to the differences in language complexity and domain spread between the corpora, but also motivates a further inspection of PHOENIX.

3. Methods

In this study, we proceed with the following research questions:

General quality issues of the corpus. Whereas several quality issues have been highlighted in previous work, we aim to obtain systematic quantitative evidence about the quality issues of the corpus. Because the German text derives from broadcast subtitles/ASR text rather than a direct translation of the signing, we treat it as weak supervision and perform an **error annotation of the training set** based on a comparison of the glosses with the original German transcription (Sections 4.1 and 5.1). We focus on the training partition, since our goal is to identify corpus-level quality issues rather than properties specific to the test set alone. To keep the annotation effort manageable while still covering the corpus broadly, we perform the annotation on a structured sample. The annotation is inspired by the Translation Error Rate (TER; Snover et al., 2009, 2006) and uses the following error categories:

- information missing in the glosses,
- information missing in the German text,
- lexical errors, and
- minor differences that do not significantly affect translation adequacy.

The "reordering" error label of TER was excluded due to the structural difference in word order between spoken German and DGS.

Quality of the test set. Given the central role of test set quality in automatic evaluation, we additionally examine to what extent the signed content is understandable to a fluent DGS signer working independently of the corpus annotations. For this purpose, they produced a back translation of the test set by watching the videos without access to either the glosses or the German transcribed text (Sections 4.2 and 5.2). The aim is to approximate what a competent human viewer can recover from the signed videos alone, which is also the basic expectation placed on automatic systems evaluated on this material. To obtain an additional indicator of interpretability, the signer recorded their translation confidence for each item. Obvious technical problems in the video material were also flagged, since such issues may directly affect comprehensibility.

Divergence between back translation and German transcription under automatic metrics. We want to assess semantic preservation under the assumed conditions of noise and weak parallelism. Additionally, we want to know to what extent the automatic metrics can

measure this semantic preservation. We therefore perform **automatic scoring of the human translations using the original German text as a reference** (Sections 4.3 and 5.3). The automatic scoring is performed with the lexical state-of-the-art machine translation automatic metric BLEU (Papineni et al., 2002), commonly reported in SLT experiments. Additionally, we report ChrF++ (Popović, 2017), TER (Snover et al., 2009), and BLEURT (Zan et al., 2024) scores.

Effect of sign-to-text back translation on scoring existing systems. We conduct an additional evaluation of previously reported SLT system outputs under different reference conditions. While the previous question focuses on how a human signer's back translations compare to the official PHOENIX references, the present experiment asks a complementary question: whether comparative system evaluation remains stable when uncertain evaluation items are removed, and the reference side is grounded more directly on the signed videos.

4. Experiment Setup

For this evaluation, both the test set and the training set were manually analysed by a human annotator through error annotation and back translation, respectively.[4]

4.1. Error Annotation of the Training Set

The training set of PHOENIX consists of 7096 pairs of gloss sequences and corresponding German sentences. A structured sample of 307 sentence pairs was annotated by one deaf fluent L2 DGS signer, requiring 25 person-hours in total.

The annotated sentence pairs were selected by starting at the beginning of the training set and sampling blocks of ten to eleven consecutive items at regular intervals, first in steps of 100 up to sentence 1500, and then in steps of 500 until the end of the corpus. This sampling strategy was chosen to obtain a structured overview of gloss and text quality throughout the full training set while covering material associated with each of the nine signers.

4.2. Sign-to-Text Back Translation of the Test Set

The test set consists of 642 video files of nine sign language interpreters. As part of our analysis, all of these files were back translated, either fully or partially, by one deaf fluent L2 DGS signer. The signer spent 55 person-hours on this task. For each item, they recorded a translation confidence score of 0, 0.5, or 1, where 1 indicates high confidence, 0.5 indicates partial or medium confidence, and 0 indicates that no reliable back translation could be produced. In addition, each video was marked for the presence or absence of noticeable technical quality problems using a binary flag (0, 1).

4.3. Scoring of Human Translation

We compute the metrics BLEU, chrF++ and TER to score the human back translation against the original German text reference. Scores were computed with SACREBLEU (Post, 2018). We repeat the scoring for 3 different versions of the back translation: translations for which the signer reported high confidence, translations for which they reported at least medium confidence, and the full set of back translations (including cases that could only be translated incompletely or not at all).

4.4. Benchmark Evaluation with Original and Sign-to-Text References

We evaluated one gloss-based and three gloss-free systems spanning a broader range of quality. The evaluated systems were TwoStream-SLT (Chen et al., 2022), SpaMo (Hwang et al., 2025), SEM-SLT (Hamidullah et al., 2024), and Signformer (Yang, 2024). The original PHOENIX test set scores serve as a baseline over the full test set of 642 items.

In addition, we evaluated the subset of 462 items for which a back translation with confidence score 1 was available. For this high-confidence subset, system outputs were scored under three reference conditions: using the original PHOENIX references only, using the back-translated references only, and

[4]The material is publicly available at https://github.com/DFKI-SignLanguage/sacre-bird-phoenix

Error type	sentences	percentage
Information missing in glosses	97	31.6%
Information missing in text	13	4.2%
Lexical errors	33	10.7%
Minor differences	165	53.7%
Total annotated parallel sentences	307	100%

Table 1: Error annotation of the training set

using both jointly in a two-reference setting.[5] This design allows us to separate three effects: the difference between the full benchmark and its subset having uncertain evaluation items removed, the effect of replacing the official reference with a sign-to-text back translation on the same items, and the effect of allowing both reference formulations simultaneously.

Before scoring, outputs and references were normalised in the same way in order to reduce superficial mismatches caused by formatting conventions. Specifically, all text was lowercased, a full stop was added at the end of each sentence, and numbers were written out in words following the conventions of the official PHOENIX German references. Automatic evaluation was then carried out with BLEU for all conditions and with BLEURT for the single-reference conditions. We focus on BLEU because it is the dominant metric in published PHOENIX-based work and therefore most directly reflects prevailing benchmark practice. BLEURT is included as a complementary semantics-oriented metric that is less dependent on exact lexical overlap. The two-reference condition was evaluated with BLEU only.

5. Results

We report findings from the corpus audit, including the manual annotation and back translation, automatic scoring, and benchmark results under different reference conditions.

5.1. General Quality Issues of the Corpus

The results of the error annotation on the training set can be seen in Table 1. Out of the 307 annotated pairs of gloss sequences and corresponding German text, substantial quality issues were found in a considerable proportion of them. The glosses were missing information as opposed to what was provided in the reference text in 31.6% of the pairs. In 4.2%, information present in the glosses was missing from the written German text. Lexical errors, such as mistranslations, could be found in 10.7% of the pairs, some of them as significant as different numbers or months. Only 53.7% of the pairs contained none or only minor differences that did not meaningfully affect translation adequacy.

The lexical errors varied in severity: some involved relatively minor meaning differences, such as good vs. exciting, clouded vs. heavy clouds, evening vs. night, rain vs. storm. Others were more consequential, including mismatches in numbers or place references, e.g. mostly vs. partially, seven vs. minus seven, North vs. Baltic Sea, and February vs. August[6].

Quality issues did not appear to be evenly distributed across the sampled material. Some sampled blocks associated with particular signer material appeared more affected than others, especially in the handling of locations, indices, and references to previous sentences. Among our 27 sampled blocks, there was only one annotated block of eleven consecutive pairs without considerable issues.

5.2. Back Translation of the Test Set

The back translation statistics for the test set are shown in Table 2. Out of the 642 video clips of the test set, 72% were back translated

[5]The two-reference setting was done only for BLEU as BLEURT does not support multiple references.

[6]There are signs for February and August that are identical in hand form and movement, but have different mouthings. Still, in PHOENIX, glosses and reference text refer to different months.

Back translation status	sentences	percentage
Successful	462	72.0%
Issues due to video quality	78	11.2%
Not translated because of lexical issues	17	2.6%
Incomplete due to unclear location signs	26	4.1%
Total number of videos	642	100%

Table 2: Statistics on the back translation of the test set

with high confidence by the annotator. 11.2% of the videos were found to have at least one noticeable quality problem, such as signs being cut off at the beginning or end, poor video quality, or the signer's face or hand movements not being visible due to overlapping text.

The translator also reported difficulties in the back translation, such as the frequently unclear use of region-specific location signs in 4.1% of the videos, and dialectal differences between signers. An example of this was "Donnerstag" (Thursday) being signed differently by Signer01 and Signer05, and Signer03 and Signer05 using different signs for "Schnee" (snow). While dialectal and regional variation is a normal and expected property of DGS rather than an error or deficiency, this can affect the suitability of PHOENIX as a benchmark.

5.3. Automatic Scores of Back Translation against Original Text

The automatic evaluation scores of the back translation scored against the original German text can be seen in Table 3. Notably, scores for all 3 automatic metrics are at the very low-end of the scoring scale (max. 14.8 BLEU), considering that an identical text scores 100 BLEU and that good quality text translations in standard test sets go up to approximately 45 BLEU.

These very low scores indicate that a substantial amount of information is lost during the full translation pipeline. As mentioned, the German text is an automatic and manually corrected transcription of what is said by the speaking TV presenter, which is then interpreted in DGS, which is then edited in the corpus segments, which is then translated by our DGS signer. Even allowing for possible back translation errors by the DGS signer when operating at high confidence, there is likely a sequence of cumulative processing and translationese effects, including:

- possible errors introduced by an early-stage open source ASR system and not fully resolved during manual correction by native German speakers,
- considerable information loss during the interpretation process, as documented by the original authors (high time pressure),
- considerable information loss during the preprocessing/segmenting of the videos and their alignment with the German transcript, and
- considerable paraphrasing that occurred during the interpretation and the back translation process, which cannot be captured by lexical metrics.

Taken together, these factors point to what is perhaps the most important conclusion of this paper, that the practice of hundreds of papers to evaluate against this particular test set using solely lexical metrics should be reconsidered.

It is also worth noting that while the high-confidence human back translation scored 14.8 BLEU, the scores reported for automatic SLT systems are in the range of 17 to 30 BLEU (e.g. Table 4 column PH14T-subset). A superficial reading could misleadingly suggest that the automatic systems outperform the human back translation. However, this is an artefact of the evaluation protocol, as automatic systems may mimic the language style and punctuation of the German text (as seen in parts of our manual analysis), or may simply obtain more coincidental matches with the reference. We may also note that the effort of the translation model developers to increase the BLEU scores of their models, by iteratively

human translation type	BLEU	chrF++	TER	BLEURT
full back translation	11.8	37.8	75.1	0.493
mid-confidence back translation	12.9	40.8	72.6	0.529
high-confidence back translation	14.8	43.6	70.4	0.573

Table 3: The human back translation of PHOENIX scored as hypotheses, using the official test set as a reference. Items below the confidence threshold are removed from both hypothesis and reference.

using PHOENIX as a development set, may result in not only learning how to translate from sign language to text, but also reproducing translation artefacts that are not relevant to translation quality.

As it can capture non-trivial semantic similarities, BLEURT appears to be less misleading than lexical-based metrics. However, it also produces scores indicating that automatic SLT systems perform equally well or better than human back translations.

5.4. SLT Benchmark Results across Reference Conditions

Results are shown in Table 4. First, the confidence=1 subset is systematically easier than the full PHOENIX test set. When evaluation is restricted to the 462 items with confidence=1 back translations while retaining the original PHOENIX references, all systems obtain higher BLEU and BLEURT than on the full 642-item benchmark. This is consistent with the fact that the subset excludes test items that a human annotator could not translate with high confidence and that SLT systems would likely also struggle to translate reliably. The subset differs in difficulty from the full test set and thus, the effects of subset selection and reference condition must be distinguished.

When the same 462 items are scored against the back-translated references instead of the original PHOENIX references, BLEU drops sharply for all systems, whereas BLEURT decreases far less. For example, TwoStream-SLT drops from 30.26 BLEU on the original reference subset to 13.39 BLEU on the back-translated reference subset, while BLEURT declines only from 0.626 to 0.587. The same pattern holds for the other systems. This suggests that the change in reference condition has a much stronger effect on lexical overlap than on semantic similarity. Put differently, system outputs remain substantially closer in meaning to the back translations than the BLEU drop alone would suggest. In addition, the confidence intervals are generally narrower under the back translation condition, most clearly for BLEU, suggesting more consistent score estimates across the evaluated items.

Overall, the confidence intervals of BLEU are so broad that no confident distinction can be made between some of the middle-ranked systems, whereas the confidence intervals of BLEURT are much narrower, indicating that relative system comparisons (ranking) made using this metric are valid. Additionally, system ranking based on BLEU is not fully stable across reference conditions. If one ignores the lack of statistical significance, on the original PHOENIX benchmark, and also on the confidence=1 subset scored with the original references, the scores suggest that SEM-SLT outperforms SpaMo in BLEU, whereas the order gets reversed once the same subset is scored against the back-translated references. Meanwhile, BLEURT appears to be more consistent across different settings and produces the same rankings as BLEU evaluated against back-translated references.

Finally, the two-reference condition yields the highest BLEU scores for all systems on the confidence=1 subset. This suggests that the back translations capture additional acceptable target realizations that are not adequately rewarded under a single-reference protocol. When both the official PHOENIX references and the back translations are treated as valid references, systems receive credit for outputs that would otherwise be penalised. This supports the interpretation that current PHOENIX evaluation is constrained both by the narrow-

Model	BLEU				BLEURT		
	PH14T full	PH14T subset	Back Transl. subset	Multi-ref subset	PH14T full	PH14T subset	Back Transl. subset
TwoStream-SLT (gloss-based)	28.2 ±2.1	30.26 ±2.39	13.39 ±1.12	34.89 ±2.22	0.597 ±0.013	0.626 ±0.010	0.587 ±0.010
SpaMo	22.2 ±2.0	24.87 ±2.64	10.59 ±1.43	27.91 ±3.04	0.542 ±0.015	0.574 ±0.013	0.543 ±0.013
SEM-SLT	23.7 ±2.1	27.27 ±2.32	8.89 ±1.47	31.01 ±2.39	0.484 ±0.012	0.511 ±0.016	0.448 ±0.013
Signformer (S3D features)	14.8 ±1.4	17.33 ±2.11	6.72 ±1.06	19.60 ±2.01	0.425 ±0.016	0.452 ±0.018	0.435 ±0.016

Table 4: Benchmark results across reference conditions. *PH14T full* denotes the full original PHOENIX test set (642 samples). All *subset* columns refer to the same 462-sample subset consisting of items for which a back translation with confidence=1 was available. *Multi-ref* uses both original and back-translated references.

ness of the reference side and by the sensitivity of lexical metrics to the exact wording of a single reference. However, the confidence intervals still do not allow a significant comparison between the two middle systems, and the observed order is different from the one suggested by the more robust BLEURT.

Taken together, these findings strengthen our main argument. The earlier human back translation experiment already showed that high-confidence human back translations receive unexpectedly low lexical scores when compared against the official PHOENIX references. The present system-level experiment extends that observation: benchmark results shift when systems are evaluated against back translations rather than against the original PHOENIX references, and they shift again when both are available. This suggests that current PHOENIX-based evaluation does not simply reflect how well systems recover the content of the signed videos, but is also influenced by the fact that the official references are tied to the original spoken-language source rather than to a direct translation of the signed material. Secondly, the lack of robustness of BLEU as a lexical metric, as compared to the semantics-aware BLEURT, is emphasised.

6. Conclusion

Our audit of the RWTH-PHOENIX-2014T corpus reveals substantial quality issues in both its training and test partitions, with systematic information loss, lexical inconsistencies, and signer-dependent variability that compromise its suitability as a benchmark for sign language recognition and translation. The low automatic scores obtained when comparing human sign-to-text back translations to the official speech-transcribed test references reveal that the test set suffers from strong translationese effects and other pipeline artefacts, and demonstrate that it does not provide a stable or linguistically faithful target for automatically evaluating systems, particularly when such evaluations rely solely on lexical metrics. The drop in scores observed when replacing the original speech-transcribed references with the human sign-to-text back translation indicates a lack of robustness of the lexical metrics and underlines that current PHOENIX-based evaluations may be too tied to the original spoken-language text, becoming unreliable in the presence of paraphrasing or reference errors.

Given the widespread use of the dataset and the risk of misleading system comparisons, we argue that future research should adopt more robust evaluation practices, incorporate human-centric and context-aware assessments, and prioritise datasets that better reflect natural sign language usage. We hope this work encourages the community to critically reassess the role of PHOENIX and to invest in the development of transparent, high-quality resources that support reliable scientific progress.

7. Acknowledgements

The research reported in this paper was supported by BMBF (now BMFTR, German Federal Ministry of Education and Research) through projects SocialWear (grant number 01IW20002) and BIGEKO (grant number 16SV9093), and by the European Union via the project SignReality, as part of financial support to third parties by the UTTER project (Horizon Europe, GA: 101070631).

8. Limitations

It should be acknowledged that while the back translations were produced by one fluent L2 signer, a stronger design would involve multiple native DGS signers in order to assess the robustness of the back translations. Also, our benchmark experiments cover only a small but deliberately heterogeneous set of systems, meaning that the observed ranking shifts should be read as evidence of sensitivity rather than as an exhaustive account of benchmark instability.

9. Ethical Considerations

We would like to note that this paper is by no means intended to be a criticism of the work of the original authors of the PHOENIX corpus, or the output produced by the interpreters. We strongly believe that the corpus has been a valuable resource for the community and has made an important contribution to scientific progress in our field. Additionally, we believe that live TV interpretation is necessary and valuable for the sign language community in terms of access to information, and the level of quality achievable under live on-air time pressure is widely accepted as fit for purpose. Here, we are focusing on the suitability of this content as part of strict evaluation protocols.

10. Bibliographical References

Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. Using Neural Machine Translation Methods for Sign Language Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284, Dublin, Ireland. Association for Computational Linguistics.

Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Eleftherios Avramidis, Vera Czehmann, Fabian Deckert, Lorenz Hufe, Aljoscha Lipski, Yuni Amaloa Quintero Villalobos, Tae Kwon Rhee, Mengqian Shi, Lennart Stölting, Fabrizio Nunnari, and Sebastian Möller. 2025. The TUB sign language corpus collection. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA. ACM.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. The ATIS sign language corpus. In *Proceedings of LREC 2008*, pages 2943–2946, Marrakech, Morocco. ELRA.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh

Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems 35, (NeurIPS 2022)*, volume 35, pages 17043–17056. Curran Associates, Inc.

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, Marseille, France. European Language Resources Association.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3785–3789, Istanbul, Turkey. European Language Resources Association (ELRA).

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. In *Proceedings of EMNLP 2020*, pages 61–71, Online. Association for Computational Linguistics.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in Size and Depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer, New York.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of IJCAI 1995*, pages 1137–1143, Montreal, Canada.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset

Compendium: Creating an Overview of Digital Linguistic Resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.

Frank J. Landy, Janet L. Barnes, and Kevin R. Murphy. 1978. Correlates of perceived fairness and accuracy of performance evaluation. *Journal of Applied psychology*, 63(6):751.

Yann LeCun, Koray Kavukcuoglu, and Clement Farabet. 2010. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of ACL 2020*, pages 4984–4997, Online. Association for Computational Linguistics.

Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. Dicta-Sign – building a multilingual sign language corpus. In *5th Workshop on the Representation and Processing of Sign Languages (LREC 2012 Satellite Workshop)*, pages 117–123, Istanbul, Turkey.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. 2024. DGS-Fabeln-1: A Multi-Angle Parallel Corpus of Fairy Tales between German Sign Language and German Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italia. ELRA and ICCL.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL 2003*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of EMNLP 2020*, pages 2685–2702, Online. Association for Computational Linguistics.

Marc Schulder, Thomas Hanke, and Maria Kopf. 2025. Making Sign Language Research Findable: The sign-lang@LREC Anthology and the Sign Language Dataset Compendium. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 277–288, Naples, Italy. Unior Press.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human

Annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA. International Association for Machine Translation.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Daniel Stein, Jens Forster, Uwe Zelle, Philippe Dreuw, and Hermann Ney. 2010. RWTH-Phoenix: Analysis of the German Sign Language Weather Forecast Corpus. In *SignLang@ LREC 2010*, pages 225–230. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Eta Yang. 2024. Signformer is all you need: Towards Edge AI for Sign Language. Version Number: 1.

Shakib Yazdani, Yasser Hamidullah, Cristina España-Bonet, Eleftherios Avramidis, and Josef van Genabith. 2026. A critical study of automatic evaluation in sign language translation. In *Proceedings of the 15th Edition of the Language Resources and Evaluation Conference (LREC-2026)*, volume -. European Language Resources Association.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building Accurate Translation-Tailored LLMs with Language Aware Instruction Tuning.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

11. Language Resource References

von Agris, Ulrich and Kraiss, Karl-Friedrich. 2010. *SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition*. ELRA, ISLRN 083-847-814-698-3.