SignBLEU: Automatic Evaluation of Multi-channel Sign Language Translation

Jung-Ho Kim*, **Mathew Huerta-Enochian***, **Changyong Ko**, **Du Hui Lee**†
EQ4ALL
Nonhyeon-ro 76-gil 11, Gangnam-gu, Seoul, Republic of Korea
{stuartkim, mathew, ericko, scottlee†}@eq4all.co.kr

Abstract

Sign languages are multi-channel languages that communicate information through not just the hands (manual signals) but also facial expressions and upper body movements (non-manual signals). However, since automatic sign language translation is usually performed by generating a single sequence of glosses, researchers eschew non-manual and co-occurring manual signals in favor of a simplified list of manual glosses. This can lead to significant information loss and ambiguity. In this paper, we introduce a new task named multi-channel sign language translation (MCSLT) and present a novel metric, `SignBLEU`, designed to capture multiple signal channels. We validated `SignBLEU` on a system-level task using three sign language corpora with varied linguistic structures and transcription methodologies and examined its correlation with human judgment through two segment-level tasks. We found that `SignBLEU` consistently correlates better with human judgment than competing metrics. To facilitate further MCSLT research, we report benchmark scores for the three sign language corpora and release the source code for `SignBLEU` at https://github.com/eq4all-projects/SignBLEU.

Keywords: evaluation metric, multi-channel language, sign language, sign language translation

1. Introduction

Sign language translation (SLT) is an emerging field that aims to bridge the gap between the Deaf, hard-of-hearing, and hearing communities. With the introduction of neural machine translation, SLT has experienced significant advancements (Camgöz et al., 2018), and innovative strategies for generating poses and videos continue to be developed (Stoll et al., 2020; Saunders et al., 2022).

A common approach to text-to-sign translation is to predict glosses, semantic labels for individual signs (Müller et al., 2023). Gloss-based SLT represents signing as a single sequence of gloss tokens, standard sequence-to-sequence modeling techniques can be used, allowing researchers to leverage the capabilities of pre-trained language models (Lee et al., 2023). However, by limiting translation to a linear gloss sequence, non-manual expressions that encapsulate additional semantic and morphological aspects of sign language as well as co-occurring manual signals are omitted. Non-manual signals convey important descriptive information, often playing the role of adjectives and adverbs (Crasborn et al., 2008; Herrmann, 2013). For example, mouthings can differentiate between identical manual signals (Woll, 2001; Crasborn et al., 2008), and eyebrow and head gestures have been shown to play a pivotal role in forming negative expressions and wh-questions (Zeshan, 2004a,b). Furthermore, co-occurring asymmetrical manual signals can convey a high-level of information or other meanings not easily represented by symmetric or sequential manual signals (Sandler, 2017). Hence, excluding these signals leads to translations that are deficient in both semantic and grammatical accuracy.

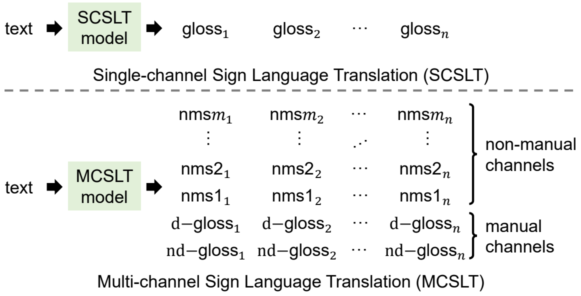


Figure 1: Comparison of SCSLT and MCSLT ("d": dominant hand, "nd": non-dominant hand, and "nms": non-manual signal).

To tackle this limitation, we introduce a new task: multi-channel SLT (MCSLT). First, we redefine the traditional gloss-based SLT that produces only a sequence of manual glosses as single-channel SLT (SCSLT). We then define MCSLT as SLT that predicts signals for *multiple channels*, allowing modeling of concurrent manual and non-manual signals (see Figure 1 for a visual comparison of the outputs of SCSLT and MCSLT). Note that predicting only two manual channels (for the dominant and non-dominant hands) simulta-

*Equal contributions.
†Corresponding author.

LREC-COLING 2024, pages 14796–14811
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

neously also qualifies as MCSLT. To our knowledge, this study is the first to specifically define and name this approach as MCSLT. We suspect that the lack of large-scale multi-channel sign language corpora and the absence of a validated metric hindered the emergence of the MCSLT task.

To facilitate meaningful development of MCSLT, we introduce SignBLEU, a new metric designed to capture both sequential and concurrent signals produced by MCSLT. We tested the proposed metric at the system level by simulating corpus translations and analyzing correlation between text-side BLEU scores and sign-side SignBLEU and other automatic metric scores. SignBLEU showed higher correlation with text-side BLEU scores than other metrics commonly used in SCSLT. We also showed that at the segment level, SignBLEU has high correlation with human evaluation of translation naturalness and fidelity and of document similarity. To support future research, we offer initial benchmark MCSLT scores on three sign language corpora.

The key contributions of our paper include:

- The introduction of the multi-channel sign language translation (MCSLT) task, emphasizing the importance of modeling multiple signing channels.
- The proposal of SignBLEU, a new metric for MCSLT, designed to assess both temporal and concurrent signals.
- Comprehensive experiments that set baseline MCSLT scores for three sign language corpora and demonstrate that SignBLEU aligns with human evaluation.

2. Related Work

We examined the factors that have influenced SLT to date, including corpora, models, and evaluation metrics. We limited analysis to studies translating text to transcribed sign language expressions.

2.1. Sign Language Corpora

Initially, sign language corpora (Sutton, 2002; Neidle, 2007; Prillwitz et al., 2008; Crasborn and Zwitserlood, 2008; Johnston, 2010) were primarily constructed for linguistic analysis of sign language expressions. Therefore, the scale of corpora was relatively small, and there was a tendency to transcribe sign language expressions in as much detail as possible. This detailed transcription was achieved using multi-tier transcription tools like ELAN (Wittenburg et al., 2006) to annotate sign language expressions across multiple signing channels or by developing image-based notations specific to sign language, such as SignWriting (Sutton, 2000) and HamNoSys (Hanke, 2004). However, the introduction of the RWTH-PHOENIX-Weather 2014 T corpus (Camgöz et al., 2018) signaled the advent of deep-learning-based SLT, prompting a shift towards large-scale data construction. Figure 2 illustrates the number of annotated channels in sign language corpora published by year. The scarcity of corpora for MCSLT relative to corpora for SCSLT can be seen clearly from this figure.

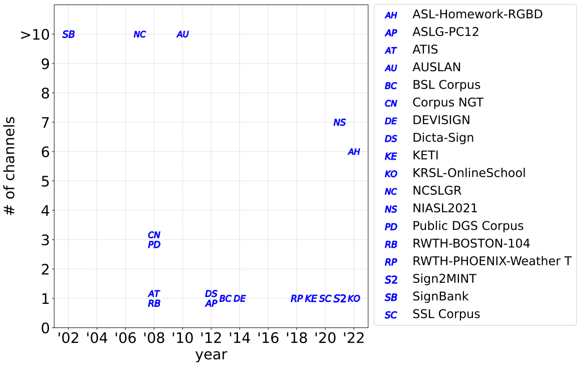


Figure 2: The number of annotated channels of published sign language corpora by year.

2.2. Sign Language Translation

Camgöz et al. (2018) introduced both an NMT-based SLT method and the RWTH-PHOENIX-Weather 2014 T corpus for SLT. SLT performance improved with the adoption of the Transformer (Vaswani et al., 2017) and with increased use of pre-trained language models as encoders (Camgöz et al., 2020; Miyazaki et al., 2020; De Coster et al., 2021). Techniques like data augmentation and multilingual NMT enhanced SLT performance (Moryossef et al., 2021; Zhu et al., 2023). However, as mentioned in §1, the above methods do not generate non-manual expressions or co-occurring expressions as they continued to be limited to predicting simplistic single-channel signals. To overcome this restriction, Jiang et al. (2023) proposed a text-to-SignWriting method and validated it by categorizing groups within the SignBank corpus (Sutton, 2002) into being either high-resource or low-resource groups. Yet, with few corpora adopting this transcription methodology, a translation approach applicable to all multi-channel sign language corpora is needed.

2.3. Evaluation Metrics for SLT

BLEU (Papineni et al., 2002) is the most widely used metric in SLT research. For reproducibility in reporting BLEU scores, many studies have recently turned to using sacreBLEU (Post, 2018). Müller et al. (2023) recommended using sacreBLEU when reporting BLEU scores in SLT and called

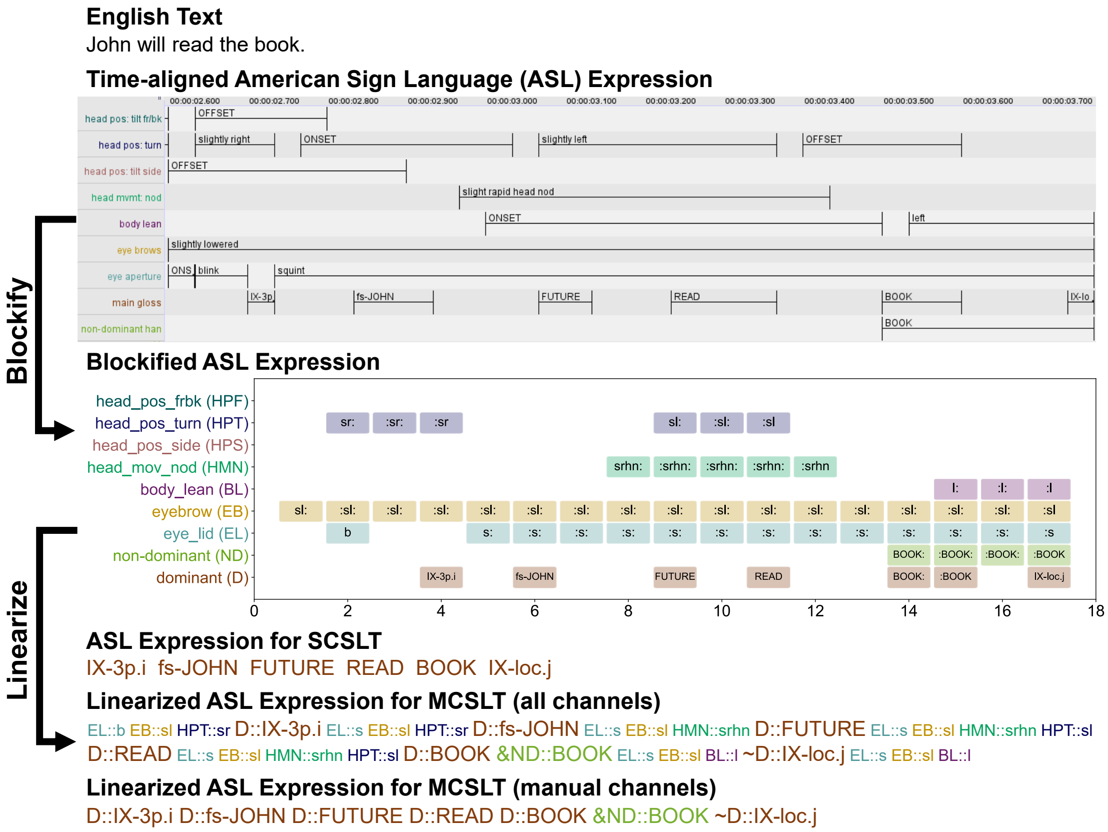

English Text
John will read the book.
Time-aligned American Sign Language (ASL) Expression
head pos: tilt fr/bk
head pos: turn
head pos: tilt side
head mvmt: nod
body lean
eye brows
eye aperture
main gloss
non-dominant han
OFFSET
slightly right
ONSET
slightly left
OFFSET
OFFSET
slight rapid head nod
ONSET
left
slightly lowered
ONS
blink
squint
IX-3p
fs-JOHN
FUTURE
READ
BOOK
IX-lo
BOOK
Blockify
Blockified ASL Expression
head_pos_frbk (HPF)
head_pos_turn (HPT)
head_pos_side (HPS)
head_mov_nod (HMN)
body_lean (BL)
eyebrow (EB)
eye_lid (EL)
non-dominant (ND)
dominant (D)
sr: :sr: :sr
sl: :sl: :sl
srhn: :srhn: :srhn: :srhn: :srhn
l: :l: :l
sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl: :sl
b
s: :s: :s: :s: :s: :s: :s: :s: :s: :s: :s: :s: :s
BOOK: :BOOK: :BOOK: :BOOK
IX-3p.i
fs-JOHN
FUTURE
READ
BOOK: :BOOK
IX-loc.j
0 2 4 6 8 10 12 14 16 18
Linearize
ASL Expression for SCSLT
IX-3p.i fs-JOHN FUTURE READ BOOK IX-loc.j
Linearized ASL Expression for MCSLT (all channels)
EL::b EB::sl HPT::sr D::IX-3p.i EL::s EB::sl HPT::sr D::fs-JOHN EL::s EB::sl HMN::srhn D::FUTURE EL::s EB::sl HMN::srhn HPT::sl D::READ EL::s EB::sl HMN::srhn HPT::sl D::BOOK &ND::BOOK EL::s EB::sl BL::l ~D::IX-loc.j EL::s EB::sl BL::l
Linearized ASL Expression for MCSLT (manual channels)
D::IX-3p.i D::fs-JOHN D::FUTURE D::READ D::BOOK &ND::BOOK ~D::IX-loc.j

Figure 3: An example of blockification and linearization.

for reporting metric signatures along with results. As BLEU is precision-based, researchers explored other types of metrics, such as ROUGE (Lin, 2004) and METEOR (Lavie and Agarwal, 2007). Researchers also employed chrF (Popović, 2015) to measure the character-level n-gram F_1 score, and TER (Snover et al., 2006) to gauge the edit distance between translated and reference sentences. We argue that any metric for MCSLT needs to be specifically adapted to handle simultaneous signals across multiple channels. Further details on our proposed method are provided in §3 and §4.

3. Towards Multi-channel SLT

To model the complex form of multiple signal channels, we introduce two transformations: blockification and linearization. Blockification converts time-aligned annotation data (e.g., ELAN's EAF format) to a unit-less sequence of co-signed blocks, and linearization converts block data to a simplified text sequence. The block representation should always be used for evaluation, while linearization allows us to apply existing sequence-to-sequence techniques to MCSLT. Figure 3 illustrates these two processes.

3.1. Blockification

ELAN and equivalent annotation formats can accurately transcribe sign language expressions down to time alignment and categorical signal attributes, but this representation is too rich to model effectively. Instead, we discretize this representation into a two-dimensional grid of equal-sized gloss blocks.

The "blockification" process can be performed in three steps. First, an ordered set of signing channels is identified along with a surjective mapping from annotation tiers to channels (e.g., multiple mouth gesture tiers not containing overlapping annotations may be mapped to a single "mouth" channel). Second, the signing timeline is segmented into maximal segments of uninterrupted signing such that barriers between segments correspond to the start or end of at least one signal and no annotation starts or ends within a segment. Third, every non-empty segment is converted to a list of gloss values, sorted according to the given tier-channel mapping and subsequent channel order.

Note that to generate information-rich blocks, we combine annotation tiers by articulator wherever possible when blockifying data. For example, annotations for "head shake" and for "head nod" can usually be combined into a single "head" channel, assuming the gestures never co-occur.

We can formally define the block representation as a gloss-valued $c' \times t'$ matrix:

$$\mathbf{B}_{c' \times t'} = \begin{bmatrix} b_{1,1} & \dots & b_{1,t'} \\ \vdots & \ddots & \vdots \\ b_{c',1} & \dots & b_{c',t'} \end{bmatrix},$$

where c' is the number of channels, t' is the number of signal overlap segments, and b_{ij} is either a gloss or $null$. The block representation is designed to capture dependent relationships across channels, so that each column is an uninterrupted segment of signing.

As a trivial example, consider the following double-channel data with gloss annotations g_1, g_2, and g_3 on channels ch_1 and ch_2:

ch_1: |———g_1———|
ch_2: |–g_2–| |——g_3——|

The block representation would then be the following 2×5 matrix:

$$\mathbf{B} = \begin{array}{r} ch_1: \\ ch_2: \end{array} \begin{pmatrix} g_1: & :g_1: & :g_1: & :g_1 & null \\ null & g_2 & null & g_3: & :g_3 \end{pmatrix},$$

where colons denote that a signal is continued to the adjacent column. See the "Blockify" transformation in Figure 3 and Appendix A.1 for more information.

Although this process removes duration information, it facilitates the modeling of sign overlap and alignment, which is a more realistic modeling target to progress to from linear gloss sequence prediction. We consider this representation to be the gold-standard for basic MCSLT and always calculate `SignBLEU` from block data (see §4.1 for more details).

3.2. Linearization

Inspired by the graph linearization technique of Bevilacqua et al. (2021), we transform block sign language expressions into a format that is compatible with existing translation models. Since manual channels typically convey the most meaning, to linearize time-aligned or block data, we first list all manual signals, ordered by signal start time as shown in the "Linearized ASL Expression for MCSLT (manual channels)" section of Figure 3. We then prefix each signal with a "D::", "ND::", or "B::" to indicate channels for the dominant hand, non-dominant hand, and both hands for two-handed signs, respectively. To model manual signal overlap, we use the following two prefix tokens:

- $\sim$: This token indicates that the current signal starts after but overlaps with the previous manual signal.
- **&**: This token indicates that the current signal starts at the same time as the previous manual signal (though their endings may differ).

Finally, we connect manual signals to co-occurring non-manuals by inserting tokens for non-manual signals directly before or after each manual token with which they overlap, as illustrated in the "Linearized ASL Expression for MCSLT (all channels)" portion of Figure 3. By convention, we interpret a sequence of identical non-manual tokens associated with adjacent manual tokens as one continued non-manual signal, though repetition and continuation can be explicitly modeled by introducing additional special tokens.

Since linearization removes signal start and end alignment, it facilitates easier application of end-to-end MCSLT and allows us to leverage large language models (LLMs) trained on text data from the same cultural region as our target sign language. Note that we consider the additional information loss in linearization to be an artifact or our current translation technology and not intrinsic to our proposed metric or task.

4. SignBLEU

In this section, we formally define `SignBLEU` as a generalization of `BLEU` to multi-channel block data.

4.1. Multi-channel N-grams

To calculate `SignBLEU`, we convert block data into n-grams using both the column and row dimensions. We propose using two types of n-grams: *temporal grams* for capturing sequential relationships within each signing channel (calculated along rows) and *channel grams* for capturing co-occurring relationships between articulators (calculated within columns). We denote the order of temporal grams and channel grams by prepending the gram length with "`t`" and "`c`", respectively. E.g., the order of temporal grams of length four is `t4`.

When calculating temporal grams, continued glosses should be seen as a single element and $null$-valued glosses should be skipped. Channel grams can be calculated as the set of unordered n-sized subsets of each column, again skipping $null$ values.

Continuing the double-channel example from §3.1, we can calculate temporal grams of order `t1` and `t2` and channel grams of order `c2` as below.

Size	Grams
`t1`	$\{ch_1g_1\}$, $\{ch_2g_2\}$, $\{ch_2g_3\}$
`t2`	$\{ch_2g_2, ch_2g_3\}$
`c2`	$\{ch_1g_1, ch_2g_2\}$, $\{ch_1g_1, ch_2g_3\}$

A comprehensive example of the n-gram calculation is presented in Appendix A.2.

Regarding n-gram generation from the block and linear representations, it is essential to highlight the following. As mentioned in §3.1, the block representation does not capture duration information—it only represents gloss overlap and alignment. Therefore, raw signing data cannot be reconstructed from the block representation. Similarly, the linear representation cannot fully represent overlap, and conversion from linear to block representation is imperfect. Since we consider the block representation to be the correct representation for MCSLT, reference grams should always be extracted from blockified annotation data, even if hypotheses are generated from linear predictions. If linearized reference data is lifted to a block representation and used to generate n-grams, modeling limitations (such as the information loss from linearization) will be ignored and `SignBLEU` scores will be inflated).

4.2. Scoring

After generating temporal and channel grams as above, `SignBLEU` calculation is analogous to the scoring method for `BLEU`, with minor adjustments.

First, modified precision is calculated for every n-gram type and order, up to the maximum order:

$$p_n^k = \frac{\sum\limits_{h \in \mathcal{H}} \sum\limits_{g \in gram_n^k(h)} Count_{clip}(g)}{\sum\limits_{h' \in \mathcal{H}} \sum\limits_{g' \in gram_n^k(h')} Count(g')},$$

where $k \in \{t, c\}$, p_n^k is the precision score of order n and gram type k, and $gram_n^k$ is the collection of all n-grams of order n and gram type k. Next, a brevity penalty is calculated to penalize short hypotheses.

$$BP = \begin{cases} 1 & \text{if } |h| > |r| \\ e^{(1-|r|/|h|)} & \text{if } |h| \leq |r| \end{cases},$$

where $|h|$ is the number of annotations in the hypothesis and $|r|$ is the number of annotations in the reference with the most similar length. Note that we use the raw annotation gloss count (not the block count) to calculate the brevity penalty (e.g, the number of glosses in the toy example from §3.1 is three). Finally, a composite score is created:

$$\texttt{SignBLEU} = BP \times e^{\left(\sum\limits_{n=1}^{n_t} w_n^t \log p_n^t + \sum\limits_{m=2}^{m_c} w_m^c \log p_m^c\right)},$$

where w_n^t is the weight for the temporal gram precision score of order `tn`, w_m^c is the weight for the channel gram precision score of order `cm`, and n_t and n_c are the maximal orders for temporal and channel grams, respectively.

To demonstrate the characteristics of different gram orders, we calculate scores for temporal orders (`t1..t4`) and channel orders (`c2..c4`) for all experiments. Since `SignBLEU` uses two gram orders, we limit each at four and report up to sixteen order-based variants in our experiments. Similar to `BLEU`, optimal gram orders and parameter values will depend on the target data and task. Note that we denote `SignBLEU` with maximal temporal order n_t and channel order n_c as SB-tn_tcn_c (e.g., for temporal and channel order 1, we write `SB-t1c1`).

`SignBLEU` can be calculated over all semantically meaningful channels. However, a manual-only variant of `SignBLEU` where n-grams are extracted only from the manual channels may be appropriate, depending on the target task and data. For reproducibility, `SignBLEU` also provides a signature, similar to `sacreBLEU` (Post, 2018) (see §5.2). A detailed scoring example is provided in Appendix A.3.

5. Experimental Settings

This section provides an overview of data, metrics, and implementations used in our experiments.

5.1. Datasets

Due to significant variation across sign language and annotation methodologies, it is crucial to assess the proposed `SignBLEU` on various datasets. To this end, we have selected three sign language corpora. Table 1 contains key statistics of each corpus. Further details on data splits and preprocessing are provided at https://github.com/eq4all-projects/SignBLEU/tree/main/reproducibility.

5.1.1. The Public DGS Corpus

The Public DGS Corpus (PDC) is part of the DGS-Korpus project, first introduced by Prillwitz et al. (2008). While PDC was not designed as a parallel corpus for training machine translation models, it features comprehensive multi-channel annotations—each hand may be annotated individually, and both mouthings and mouth gestures have annotations—coupled with aligned German and English sentences. We employ the third release (Konrad et al., 2020), which incorporates the most recent updates as of 2020.

5.1.2. NIASL2021

Huerta-Enochian et al. (2022) introduced the NIASL2021 corpus (NS21), a large-scale Korean-Korean Sign Language (KSL) parallel corpus for SLT, in 2021. The corpus, based on emergency alerts and weather forecasts, includes non-manual

		PDC	NS21	NCSLGR
Language Pair		German-DGS	Korean-KSL	English-ASL
# Instances	train	61,912	29,980	1,124
	dev.	983	1,397	375
	test	985	1,398	375
Annotated Channels		hands, mouth	hands, head, eyebrows, cheeks, mouth	hands, head, eyebrows, eyes, mouth
Vocabulary Size	source	19,947	4,323	1,994
	target	4,674	4,503	918
Domain		deaf culture	emergency alerts	short stories

Table 1: Key statistics of sign language corpora.

signals from the head, eyebrows, cheeks, and mouth, as well as separate annotations for each manual channel.

5.1.3. NCSLGR

We use the ELAN version of Boston University's The National Center for Sign Language and Gesture Resources corpus (NCSLGR) (Neidle and Sclaroff, 2012). We use this corpus as it contains the highest number of annotation tiers, despite its relatively small size.

5.2. Metrics

To evaluate the utility of SignBLEU, we pitted it against standard metrics used in SCSLT, including BLEU, TER, chrF, METEOR, and ROUGE (specifically ROUGE-L F_1). Non-SignBLEU metrics were calculated on linearized data (see §3.2).

We calculate two sets of metrics for each experiment. First we calculate scores using all channels and then again for representations of the manual channels only. This allows us to better explore the characteristics of each metric. We provide the SignBLEU signature[1] used in our experiments and sacreBLEU version 2.3.1 signatures for BLEU[2], TER[3], and chrF[4].

5.3. Implementation Details

We fine-tuned pre-trained LLMs on each test corpora. We used BLOOM-CLP-German 1.5B model[5] for German-to-DGS, Ko-GPT-Trinity 1.2B model[6] for Korean-to-KSL, and TinyLlama 1.1B model[7] for English-to-ASL translation. Each model was fine-tuned and tested on one NVIDIA A100 80GB GPU. We present hyperparameter search results for each model in Table 2.

Hyperparameter	Search Space	Pick		
		PDC	NS21	NCSLGR
# Epochs	$\{1, \ldots, 8\}$	2	3	6
LR	$[10^{-6}, 10^{-4}]$	$3.9 * 10^{-5}$	$6.0 * 10^{-5}$	$8.4 * 10^{-5}$
Grad. accum.	$\{4, 8, 16, 32\}$	8	8	8
LoRA	$\{T, F\}$	F	F	F
Warm start	$\{T, F\}$	F	T	F
Batch size	$\{8, 16\}$	8	16	16

Table 2: Hyperparameter search results.

Channels	Metric	PDC		NS21		NCSLGR	
		Dev.	Test	Dev.	Test	Dev.	Test
All	SB-t1c1	20.15	19.43	24.75	26.27	21.70	22.02
	SB-t1c2	14.25	13.43	21.19	22.67	18.01	18.30
	SB-t1c3	0.00	0.00	16.02	17.60	12.27	12.45
	SB-t1c4	-	-	11.87	13.50	7.24	7.36
	SB-t2c1	9.76	9.05	13.49	15.15	10.51	10.06
	SB-t2c2	9.87	9.13	14.90	16.50	11.82	11.55
	SB-t2c3	0.00	0.00	13.19	14.78	9.85	9.70
	SB-t2c4	-	-	10.79	12.38	6.75	6.69
	SB-t3c1	4.89	4.20	7.52	9.17	5.88	5.06
	SB-t3c2	5.86	5.12	9.38	11.08	7.42	6.66
	SB-t3c3	0.00	0.00	9.33	10.99	7.04	6.47
	SB-t3c4	-	-	8.36	9.96	5.43	5.08
	SB-t4c1	2.50	2.08	4.22	5.82	3.15	2.55
	SB-t4c2	3.30	2.81	5.65	7.42	4.30	3.65
	SB-t4c3	0.00	0.00	6.13	7.87	4.51	3.94
	SB-t4c4	-	-	5.92	7.59	3.85	3.44
Manual	SB-t1c1	19.56	19.23	20.96	23.32	20.75	20.97
	SB-t1c2	0.00	0.00	20.14	22.33	5.52	6.73
	SB-t2c1	9.54	9.21	10.47	13.07	13.71	13.23
	SB-t2c2	0.00	0.00	12.85	15.40	6.51	7.23
	SB-t3c1	4.87	4.40	5.65	8.18	9.44	8.59
	SB-t3c2	0.00	0.00	7.69	10.40	5.93	6.08
	SB-t4c1	2.57	2.13	3.26	5.51	6.29	5.67
	SB-t4c2	0.00	0.00	4.66	7.23	4.70	4.68

Table 3: MCSLT Benchmark scores.

6. Experimental Results

We calculated and report SignBLEU benchmark scores on the test set of each sign language corpus (see §6.1). We then validated SignBLEU by analyzing its correlation with text-based BLEU scores at the system level and with human evaluation at the segment level (see §6.2 and §6.3). Finally, we developed guidelines on how to interpret SignBLEU scores and offer suggestions for its usage (see §6.4).

6.1. MCSLT Benchmark Scores

We present MCSLT benchmark scores for the test sets in Table 3. As mentioned in §1, these are *initial* MCSLT benchmarks, and we share them with

[1] off:na||t:3|c:2|dim:1||m:sbleu|sm:exp|eff:n||v:0.1.0

[2] nrefs:1|case:mixed|eff:no|tok:none|smooth:exp

[3] nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no

[4] nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no

[5] https://huggingface.co/malteos/bloom-1b5-clp-german

[6] https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5

[7] https://huggingface.co/TinyLlama/TinyLlama-1.1B-intermediate-step-480k-1T

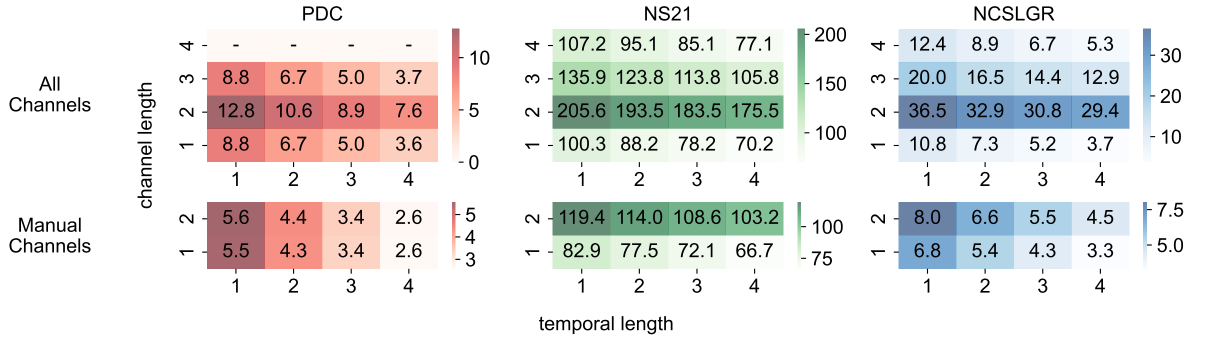


Figure 4: The average $gram_t$ and $gram_c$ counts per sentence by maximum gram order.

the hope that we can encourage further research and advancements in MCSLT.

We report all-channel and manual-channel SignBLEU scores up to gram order t4c4, resulting in 24 different metrics. Naturally, it can be challenging to determine which metrics to prioritize for each corpus. We suggest analyzing both gram frequencies and annotation methodologies as a good starting point. Figure 4 presents the average n_t-gram and n_c-gram counts per sentence categorized by temporal and channel lengths. These counts not only reveal the characteristics of each corpus, but can also provide a glimpse into which temporal and channel levels it would be beneficial to focus on for each corpus.

6.2. Correlation with Text-side BLEU

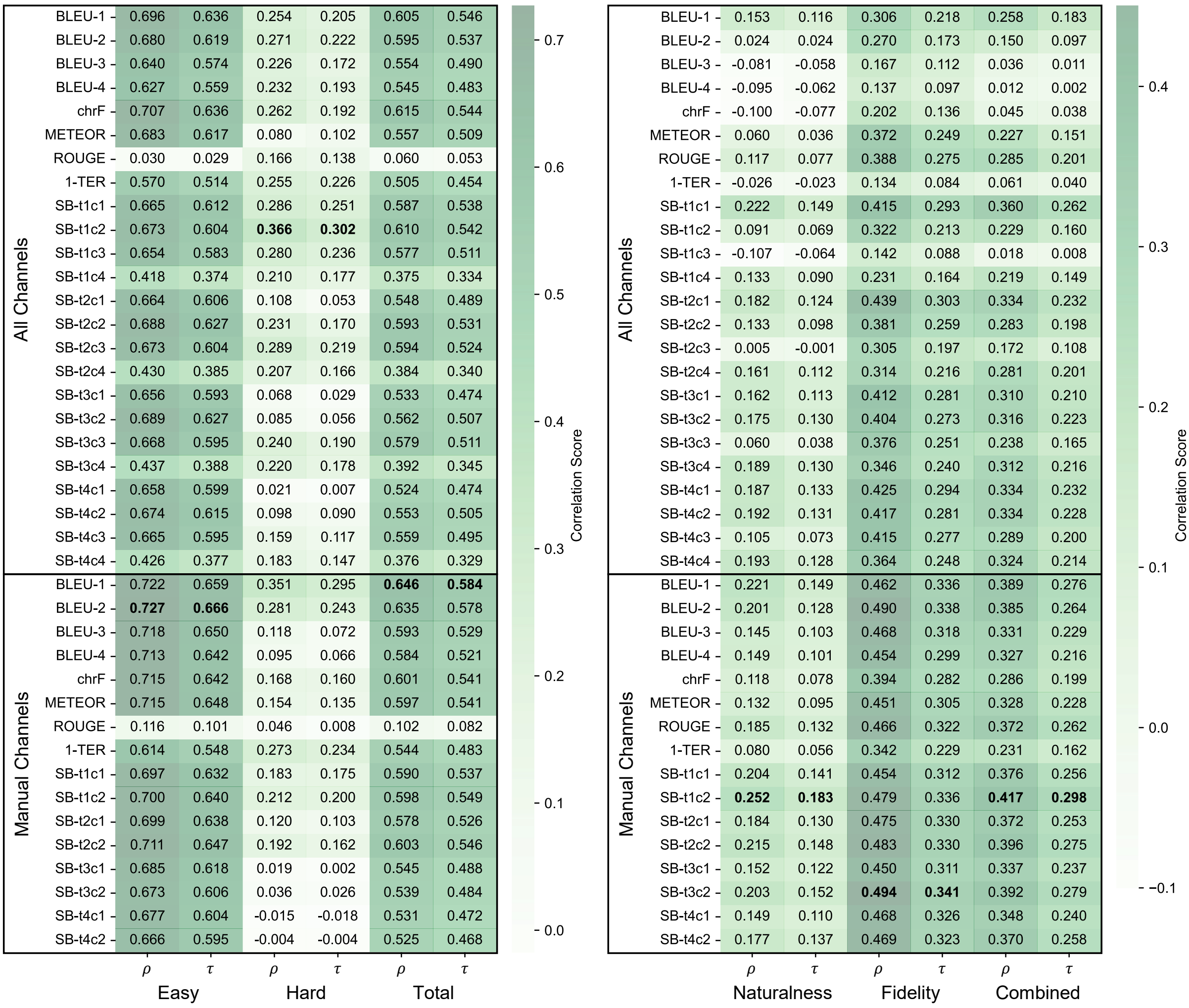
Inspired by the success of backtranslation for translation evaluation and given that our three test sets have both text-side and sign-side annotations, we measured the correlation between corpus BLEU scores on the text side and various corpus metric scores on the sign side. To do so, we make two assumptions: supplied text translations of sign language data are of high quality and corpus BLEU is reliable for system evaluation.

We simulated translation systems using randomly sampled hypothesis and reference sentences. For each corpus and simulation run, we sampled 200 instances, splitting the samples into two sets of 100 instances. We used one set as reference translations and one set as hypothesis translations. We then scored the simulated system with each metric. We repeated these sampling and scoring steps 10,000 times to get 10,000 system scores. Finally, we calculated rank correlation (using Spearman's Rho and Kendall's Tau-b) between each sign-side metric and the text-side BLEU scores. Unlike most system-level analysis, this simulation does not compare system performance over the same instances. However, due to the law of large numbers, this will not matter given enough samples.

Channels	Metric	PDC		NS21		NCSLGR	
		ρ	τ	ρ	τ	ρ	τ
All	Existing Metrics						
	BLEU-1	.206	.140	.153	.104	.128	.083
	BLEU-2	.258	.178	.228	.153	.276	.184
	BLEU-3	.264	.182	.272	.183	.450	.302
	BLEU-4	.267	.184	.251	.171	.457	.305
	chrF	.127	.085	.127	.085	.205	.136
	METEOR	.229	.154	.199	.135	.277	.180
	ROUGE	.225	.151	.196	.133	.242	.160
	1-TER	.081	.055	.099	.065	-0.021	-0.013
	SignBLEU						
	SB-t1c1	.207	.140	.186	.125	.248	.163
	SB-t1c2	.238	.169	.211	.142	.260	.174
	SB-t1c3	.054	.044	.193	.130	.325	.226
	SB-t1c4	-	-	.156	.106	.400	.321
	SB-t2c1	.229	.173	.205	.138	.389	.261
	SB-t2c2	.259	.201	.230	.154	.372	.248
	SB-t2c3	.064	.053	.217	.146	.359	.251
	SB-t2c4	-	-	.178	.120	.403	.325
	SB-t3c1	.401	.326	.189	.127	.543	.416
	SB-t3c2	.410	.334	.211	.142	.543	.414
	SB-t3c3	.064	.053	.220	.149	.537	.417
	SB-t3c4	-	-	.197	.134	.451	.367
	SB-t4c1	.390	.318	.172	.117	.648	.525
	SB-t4c2	.389	.318	.185	.126	.648	.524
	SB-t4c3	.064	.053	.201	.137	.618	.498
	SB-t4c4	-	-	.207	.142	.466	.380
Manual	Existing Metrics						
	BLEU-1	.186	.125	.221	.148	.362	.249
	BLEU-2	.264	.181	.309	.211	.604	.431
	BLEU-3	.274	.189	.305	.209	.614	.441
	BLEU-4	.279	.193	.301	.207	.613	.439
	chrF	.146	.099	.255	.173	.445	.308
	METEOR	.224	.151	.262	.178	.488	.335
	ROUGE	.208	.140	.236	.161	.372	.251
	1-TER	.081	.054	.122	.081	-0.026	-0.017
	SignBLEU						
	SB-t1c1	.191	.128	.238	.161	.369	.252
	SB-t1c2	.062	.050	.235	.160	.368	.292
	SB-t2c1	.320	.255	**.326**	**.223**	.623	.477
	SB-t2c2	.084	.069	.320	.218	.449	.364
	SB-t3c1	**.422**	**.345**	.319	.219	**.703**	**.570**
	SB-t3c2	.084	.069	.319	.218	.475	.387
	SB-t4c1	.389	.318	.280	.204	.689	.561
	SB-t4c2	.070	.057	.280	.204	.485	.395

Table 4: System level correlations of text-side BLEU with multiple sign language metrics. We highlight the **top-1** and top-5 highest correlation scores for readability.

Table 4 provides a summary of the correlation results. All three datasets showed higher correlation



(a) similarity ranking

(b) direct assessments

Figure 5: Correlations with human judgments. Highest scores are highlighted in bold for readability.

with a SignBLEU variant than with other metrics. We can also see features of each dataset reflected in the different gram order scores. For example, the difference between single-channel and double-channel manual SignBLEU correlation for PDC is likely due to how most PDC manual signs are annotated for only a single channel. Contrast this with NS21 which contains mostly two-handed (symmetric and asymmetric) manuals and shows little difference between single-channel and double-channel manual SignBLEU scores.

6.3. Correlation with Human Judgment

We performed two additional segment-level experiments to explore agreement with human evaluation.

6.3.1. Task #1: Ranking

We examined the correlation between sentence-level similarity rankings by SignBLEU and human-based similarity rankings on NS21, using a set of 147 questions. For each question, a reference sign language video and four candidate videos were provided. Most videos were between ten and twenty seconds long. Evaluators ranked the candidate videos by similarity to the reference video and were instructed to base their rankings on meaning similarity first and signing flow and vocabulary use second. Out of the 147 questions, 115 were easier to rank (termed "**Easy**") as the candidate videos were randomly sampled from both the same and different domains as the reference video, leading to larger semantic differences between candidate videos. The remaining 32 questions, labeled "**Hard**", posed a greater challenge. Their candidate videos were intentionally sampled to have a unigram gloss precision of over 90%, requiring close scrutiny to determine similarity rankings. While this is not a translation task, it provides insight into the MCSLT and SignBLEU.

Four native signers individually ranked candidate videos. We then aggregated their responses into labels, allowing for ties. Metric-based similarity was calculated with our test metrics applied to existing annotations for each video. To reduce bias, candidate videos were selected with one additional criterion–all five videos had to have either the same signer wearing the same outfit or have different signers.

Figure 5a presents the rank correlation as a heatmap. Correlation was computed using both Spearman's Rho (ρ) and Kendall's Tau-b (τ). Manual `BLEU-1` and manual `BLEU-2` showed the highest agreement with human scores on the total ranked dataset and on the "Easy" subset, respectively. Post-evaluation interviews with the four evaluators revealed that for the "Easy" ranking task, non-manuals and co-occurring signs could be completely ignored and almost all candidate videos could still be correctly ranked. Thus, it makes sense that a simple `BLEU-1` or `BLEU-2` score would perform well for this task. On the other hand, rankings from the "Hard" subset showed higher correlation with all-channel `SignBLEU` scores, especially scores of channel order `c2`, `c3`, and `c4`. This suggests that evaluation of multiple channels, including non-manual channels, was required to effectively rank the candidates.

It is worth noting that since NS21 was constructed from manually-translated emergency alerts and weather broadcasts, non-manual signals may play a secondary role to manual signals, in contrast to collected from spontaneous signing.

6.3.2. Task #2: Direct Assessments

Since our primary objective was to validate `SignBLEU` as a metric for machine translation, we collected native signer direct assessments of automatic translation results and compared them with metric scores. We randomly selected 53 instances from the development subset of NS21 and generated translations for each instance using the same model used to report NS21 benchmark scores. To avoid bias introduced by the influence of an avatar representation, we hired two experienced signers to create signing videos based on the translations by re-signing the predicted multi-channel glosses. Eighteen evaluators then scored each video. For each instance, each evaluator first watched the re-signed video and scored it for naturalness. They then viewed one of the correct reference translation videos and scored the re-signed video for fidelity. All training and evaluation was conducted in sign language. Naturalness and fidelity were both evaluated on eleven-point Likert scales labeled uniformly from 0 to 100.

Again, we analyzed correlation between evaluator and metric-based scores using Spearman's Rho (ρ) and Kendall's Tau-b (τ). Results are displayed in Figure 5b. Note that we used z-scores calculated separately over each evaluator's naturalness and fidelity scores for correlation analysis. "**Combined**" was calculated from the mean of naturalness and fidelity z-scores.

All metrics showed higher correlation with fidelity than with naturalness. This aligns with results from the "Easy" subset of the similarity ranking experiment, where metrics evaluated on manual channels demonstrated higher correlations than those on all channels. These experimental results further illustrated that NS21 is more biased towards manual information. Overall, we found that `SignBLEU` outperformed existing metrics. Interestingly, manual `SB-t1c2`, which emphasizes co-occurring signs, showed the highest correlation with human-scored naturalness, and manual `SB-t3c2`, which captures a sequential relationship in addition to co-occurring relationships, showed the best correlation with human-scored fidelity.

6.4. SignBLEU Guideline

To use `SignBLEU`, an appropriate max gram order must be selected. One can simply use the `t1c2` variant due to the high number of temporal and channel grams of this order, as seen in Figure 4. This variant also showed high correlation with human judgment on NS21. However, it showed poor and mediocre correlation with manual-only and all-channel PDC, respectively.

If human evaluation is available for one's corpus, it should be utilized to find appropriate gram orders. If it is not available, but text-side translations for your data are, we recommend performing correlation analysis with text-side `BLEU`, as in §6.2.

To help with gram order and other parameter selection, we will publish additional analysis online at https://github.com/eq4all-projects/SignBLEU.

7. Conclusions and Future Work

In this study, we proposed a new gloss-based sign language translation (SLT) task that we termed multi-channel sign language translation (MCSLT). MCSLT refers to any SLT that generates gloss predictions across multiple signal channels. We then proposed and validated a new metric, `SignBLEU`, for MCSLT evaluation. We hope that more SLT research will adopt the multi-channel approach, and we will continue to evaluate and refine `SignBLEU` as an open-source solution to MCSLT evaluation.

8. Acknowledgment

This work was partly supported by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Ministry of Science and ICT (MSIT, Korea) (No. 2022-0-00010, Development of Korean sign language translation service technology for the deaf in medical environment) and the Technology Innovation Program funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) (No. 20014406, Development of interactive sign language interpretation service based on artificial intelligence for the hearing impaired).

9. Ethical Considerations

The study's protocol was approved by the Korea National Institute for Bioethics Policy (IRB No. P01-202310-01-014). All participants were informed about the nature, purpose, procedures, potential risks, and benefits of the research. They gave their voluntary consent to participate, ensuring they felt no pressure. Moreover, they were made aware of their right to withdraw from the study at any time without any repercussions.

To ensure participant compensation, we collected certain personal information. However, upon completing the compensation-related administrative processes, all personal data was destroyed. For the sake of data security, access to the evaluation data was restricted to the authors, all of whom are registered researchers under the research plan sanctioned by the IRB.

Our study, aimed at comparing and assessing sign language videos, necessitated the inclusion of deaf individuals who use sign language as their primary language of communication. Every step, ranging from recruitment and guideline explanation to the evaluation itself, was communicated in sign language to ensure clear communication. Risks for participants were kept to a minimum.

Participants spent a maximum of 2 hours in the study, spanning the time from introduction to the evaluation guidelines through to the completion of the actual evaluation. As compensation for their time and insights, they received payment exceeding the national minimum wage.

10. Limitations

- This study was focused on developing and validating a metric for automatic evaluation of MCSLT. Although the translation model used in our experiments was optimized through hyperparameter search, the reported scores should be considered only as preliminary benchmark scores for MCSLT, and we do not consider our modeling approach itself to be a technical contribution.

- Linearized sign language expressions and equivalent multi-channel sign language expressions predicted by MCSLT models are human-readable but are not directly viewable as sign language expressions. Therefore, to conduct human evaluations of naturalness and fidelity, we presented the output as a sign language video re-signed by native signers so that the evaluator would not be negatively biased by either raw visualization or by an avatar representation. However, this approach required re-signing the predicted MCSLT *exactly*, which proved extremely difficult. While we cannot guarantee that we were able to eliminate all production bias, we conducted several rounds of review for each video to remove extraneous signals. Since most errors could be identified quickly, the bigger challenge was simply the energy- and time-cost of re-signing. Since the synthetic utterances included many small "errors", signers had to practice each utterance before filming and most videos were re-filmed at least once. Due to these costs, we advise against using this approach and emphasize the need to find a better solution to isolated human evaluation of SCSLT and MCSLT results, unbiased by avatar and other production methodologies.

- Though we provided some interpretation as to why certain max gram order variants performed well or poorly, it is important to recognize that the optimal choice of max gram order will depend on the target corpus and the user's specific objectives.

- All corpora used in this study contain different language pairs (German-DGS, Korean-KSL, and English-ASL). Due to this, there were ethical and accessibility-related limitations to performing user evaluations for every corpus. To alleviate this to some extent, and inspired by the practice of assessing quality through backtranslation in sign language production research, we calculated correlation with text side BLEU score, attempting to provide as objective a validation as possible for all sign language corpora.

11. Bibliographical References

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Onno Crasborn, Els van der Kooij, Dafydd Waters, Bencie Woll, and Johanna Mesch. 2008. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, 11(1):45–67.

Onno Crasborn and Inge Zwitserlood. 2008. The Corpus NGT: an online corpus for professionals and laymen. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 44–49, Marrakech, Morocco. European Language Resources Association (ELRA).

Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. 2021. Frozen pretrained transformers for neural sign language translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 88–97, Virtual. Association for Machine Translation in the Americas.

Thomas Hanke. 2004. HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.

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).

Annika Herrmann. 2013. *Modal and focus particles in sign languages: A cross-linguistic study*, volume 2. Walter de Gruyter.

Mathew Huerta-Enochian, Du Hui Lee, Hye Jin Myung, Kang Suk Byun, and Jun Woo Lee. 2022. KoSign sign language translation project: Introducing the NIASL2021 dataset. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 59–66, Marseille, France. European Language Resources Association.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Huije Lee, Jung-Ho Kim, Eui Jun Hwang, Jaewoo Kim, and Jong C. Park. 2023. Leveraging large language models with vocabulary sharing for sign language translation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Taro Miyazaki, Yusuke Morita, and Masanori Sano. 2020. Machine translation from spoken language to sign language using pre-trained language model as encoder. 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 139–144, Marseille, France. European Language Resources Association (ELRA).

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

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.

Carol Neidle. 2007. Signstream annotation: Addendum to conventions used for the american sign language linguistic research project.

Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: Development of the asllrp data access interface (dai).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. 2008. DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association (ELRA).

Wendy Sandler. 2017. The challenge of sign language phonology. *Annual Review of Linguistics*, 3(1):43–63.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5141–5151.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Valerie Sutton. 2000. Signwriting. *Deaf Action Committee (DAC) for Sign Writing*.

Valerie Sutton. 2002. Signbank. *Retrieved online at: https://www.signbank.org*, 2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Bencie Woll. 2001. he sign that dares to speak its name: echo phonology in british sign language (bsl). *P. Boyes Braem & R. Sutton-Spence (eds.)*, pages 87–98.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Ulrike Zeshan. 2004a. Hand, head, and face: Negative constructions in sign languages. *Linguistic Typology*.

Ulrike Zeshan. 2004b. Interrogative constructions in signed languages: Crosslinguistic perspectives. *Language*, 80(1):7–39.

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.

12. Language Resource References

Carol Neidle and Stan Sclaroff. 2012. *National Center for Sign Language and Gesture Resources (NCSLGR) corpus*. Boston University. ISLRN, American Sign Language Linguistic Research Project (ASLLRP), ISLRN 833-505-711-564-4.

Appendix A. Calculation Example

This section illustrates how `SignBLEU` is calculated using two example documents, document 1 and document 2, as seen in figure 6 (top and bottom, respectively). This appendix is provided to supplement the explanations for blockification from §3 and both `SignBLEU` scoring (including gram creation) from §4.

See Appendix A.1 the example blockification calculation and Appendix A.2 for gram calculation. Scoring is an extension of the modified n-gram precision scoring from the original `BLEU` algorithm, and the calculation for this example is covered briefly in Appendix A.3.

The two examples shown here are synthetic documents containing some degree of gloss overlap. Manual tiers are "both" for signals using both hands and "right" for right-hand only signals. All other tiers are non-manual tiers.

A.1. Multi-Channel Blocks

Blocks can be generated iteratively using annotation start and end times.

Let G denote a collection of gloss annotations; let $T = \{t_i\}$ denote the collection of all gloss start and end times, de-duplicated and sorted in ascending order; and let $g.start$, $g.end$, $g.tier$, and $g.name$ denote annotation start, end, tier, and gloss name for annotation g. Also assume that we have a mapping $M : tier \mapsto channel$ that maps tiers to target channels. M need not be injective as we may want to map multiple tiers to the same signal channel. The block representation B of a document can then be calculated using algorithm 1.

Algorithm 1 blockify$(G, T, M) : B$

```
$n \leftarrow (|T| - 1)$
$B \leftarrow \{\}$
for $i \in \{1...n\}$ do
    ▷ Initialize block dictionary
    $block \leftarrow \{\}$
    $gs \leftarrow \{g|g \in G,$
            $g.start \leq t_i < t_{i+1} \leq g.end\}$
    for $g \in gs$ do
        ▷ Denote continuation
        $prefix \leftarrow$ " : " if $g.start < t_i$ else " "
        $suffix \leftarrow$ " : " if $g.end > t_{i+1}$ else " "
        $name \leftarrow prefix + g.name + suffix$
        $channel \leftarrow M(g.tier)$
        $block[channel] \leftarrow name$
    end for
    if $|block| > 0$ then
        $B.append(block)$
    end if
end for
Return $B$
```

This generates a sequence of blocks, where each block maps channels to gloss names. By convention, we add a key-value pair for each missing channel, mapping the channel to `null`. Glosses may be renamed by pre- or post-pending a special symbol (shown here as " : ") to the gloss name to mark continuation from the previous or to the next block, respectively. Continuation identifiers are used to calculate intra-channel (temporal) grams directly from the block representation and are used in several `SignBLEU` variants that we are still developing and plan to release in the future. Given any fixed channel order γ, we can represent a block sequence as a block matrix by converting each block to a column vector with values ordered by the order of their keys in γ. We consider the block matrix synonymous with the block sequence representation and refer to them both as block representations.

See Table 5 for example block representations of both ELAN examples from Figure 6.

A.2. Temporal and Channel Grams

Given annotation data represented as a block matrix, n-grams can be easily calculated by extracting n adjacent glosses from each row across blocks (temporal grams) and sets of size n of non-`null` glosses across channels from within each column (channel grams).

A.2.1. Temporal Grams

Given a block matrix B, temporal grams can be calculated as

$$\bigcup_{row \in B} gram_n(\{b \mid b \in row, b \neq \texttt{null}, \neg pre(b)\}),$$

where $gram_n$ is the standard n-gram function and pre is true if and only if there is a continuation prefix. All experiments from this study used this simple implementation to extract temporal grams of order `t1`..`t4` from each channel. Since channels may be constructed from multiple tiers during blockification, extracting temporal grams from the block representation may be easier than from the original time-aligned annotation representation. Simply collect adjacent non-`null` glosses, skipping those that start with continuation markers.

We are experimenting with including whitespace (`null` values) and with weighting based on the number of blocks a single signal spans, and we may introduce parameters to allow for different temporal gram calculations in the future.

A.2.2. Channel Grams

Channel grams are intra-block, inter-channel grams (i.e., constructed from within a single block column). However, since channels have no inherent order, channel grams of size n from a given block are the set of all n-length subsets of non-`null`-annotations from that block. When calculating both temporal and channel grams, we skip channel grams of order `c1` since the high level of overlap between temporal grams of order `t1` and channel grams of order `c1` led to worse performance.

A.2.3. 2D Grams

We experimented with two-dimensional grams constructed from both the temporal and channel dimensions, but the combination of separate temporal and channel grams performed better than the implementations of two-dimensional grams that we tested. Two-dimensional grams also suffer from two other challenges: they are more sensitive to small alignment changes and they lead to a much higher computational complexity due to the increased number of unique grams. We plan to continue improving the two-dimensional implementation as a possible future improvement.

A.3. Scoring

As stated above, scoring is analogous to that of the original `BLEU` algorithm, with adjustments to handle multiple types of n-grams.

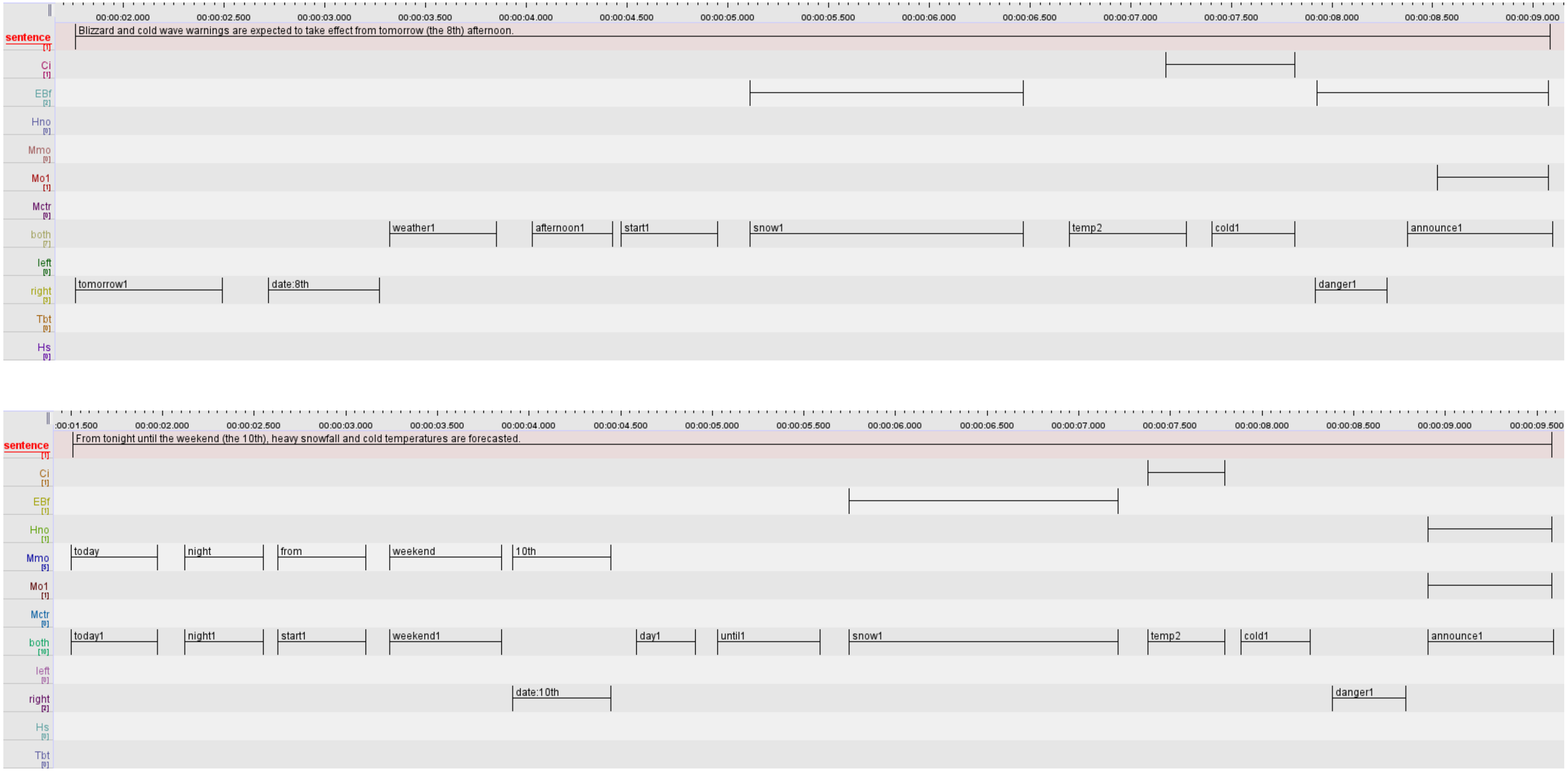
We found that weighting each gram order and type evenly performed well in our original experiments and recommend doing so as a safe starting point. We calculate the brevity penalty using the number of annotations included in the calculation, which can be calculated from the block representation by counting glosses that do not start with a continuation prefix. We tested several other variants, including the number of blocks, the number of glosses (with or without a continuation prefix), and the number of blocks containing manual glosses, but the simple annotation count performed the best in our initial experiments.

For the hypothesis and reference presented in Figure 6, the modified precision for orders `t1`, `t2`, `t3`, and `c2` are as follows:

Order	Score
`t1`	0.368421
`t2`	0.266667
`t3`	0.181818
`c2`	0.625

Finally, we can calculate the raw aggregate score, the brevity penalty, and the final `SignBLEU` score:

	Score
Raw	0.325056
BP	0.768621
`SignBLEU`	0.249844


Blizzard and cold wave warnings are expected to take effect from tomorrow (the 8th) afternoon.

From tonight until the weekend (the 10th), heavy snowfall and cold temperatures are forecasted.

Figure 6: Sample ELAN instances.

Doc	Channel	Blocks																					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	right	tomorrow1	date:8	weather1	afternoon1	start1	snow1:	:snow1	-	temp2:	:temp2	-	cold1	-	danger1:	:danger1	-	-	-				
	left	-	-	weather1	afternoon1	start1	snow1:	:snow1	-	temp2:	:temp2	-	cold1	-	-	-	-	-	-				
	eye	-	-	-	-	-	-	EBf:	:EBf	-	-	-	-	-	-	EBf:	:EBf:	:EBf	-				
	mouth	-	-	-	-	-	-	-	-	-	Ci:	:Ci:	:Ci:	:Ci	-	-	-	Mo1:	:Mo1				
2	right	-	night1:	:night1	start1:	:start1	-	-	weekend1:	:weekend1	-	date:10:	:date:10	day1	until1	snow1:	:snow1	-	temp2:	:temp2	-	cold1	danger1
	left	-	night1:	:night1	start1:	:start1	-	-	weekend1:	:weekend1	-	-	-	day1	until1	snow1:	:snow1	-	temp2:	:temp2	-	cold1	-
	eye	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	EBf:	:EBf	-	-	-	-	-
	mouth	Mmo:	:Mmo	-	-	Mmo:	:Mmo	Mmo:	:Mmo	-	Mmo:	:Mmo	-	-	-	-	-	-	-	Ci:	:Ci	-	-

Table 5: Example blocks

Order	Grams											
	Doc 1						Doc 2					
t1	eye_EBf ×2,	right_tomorrow1,	right_date:8,	right_weather1,	right_afternoon1,	right_start1,	eye_EBf,	right_night1,	right_start1,	right_weekend1,	right_date:10,	right_day1
	right_snow1,	right_temp2,	right_cold1,	right_danger1,	left_weather1,	left_afternoon1,	right_until1,	right_snow1,	right_temp2,	right_cold1,	right_danger1,	left_night1,
	left_start1,	left_snow1,	left_temp2,	left_cold1,	mouth_Ci,	mouth_Mo1	left_start1,	left_weekend1,	left_day1,	left_until1,	left_snow1,	left_temp2,
							left_cold1,	mouth_Mmo ×4,	mouth_Ci			
t2	(eye_EBf	eye_EBf),	(right_tomorrow1	right_date:8),	(right_date:8	right_weather1),	(right_night1	right_start1),	(right_start1	right_weekend1),	(right_weekend1	right_date:10),
	(right_weather1	right_afternoon1),	(right_afternoon1	right_start1),	(right_start1	right_snow1),	(right_date:10	right_day1),	(right_day1	right_until1),	(right_until1	right_snow1),
	(right_snow1	right_temp2),	(right_temp2	right_cold1),	(right_cold1	right_danger1),	(right_snow1	right_temp2),	(right_temp2	right_cold1),	(right_cold1	right_danger1),
	(left_weather1	left_afternoon1),	(left_afternoon1	left_start1),	(left_start1	left_snow1),	(left_night1	left_start1),	(left_start1	left_weekend1),	(left_weekend1	left_day1),
	(left_snow1	left_temp2),	(left_temp2	left_cold1),	(mouth_Ci	mouth_Mo1)	(left_day1	left_until1),	(left_until1	left_snow1),	(left_snow1	left_temp2)
							(left_temp2	left_cold1),	(mouth_Mmo	mouth_Mmo) ×3,	(mouth_Mmo	mouth_Ci)
c2	(left_weather1	right_weather1),	(left_afternoon1	right_afternoon1),	(left_start1	right_start1),	(left_night1	right_night1) ×4,	(mouth_Mmo	right_night1),	(left_night1	mouth_Mmo),
	(left_snow1	right_snow1) ×2,	(eye_EBf	right_snow1),	(eye_EBf	left_snow1),	(left_start1	right_start1) ×2,	(mouth_Mmo	right_start1),	(left_start1	mouth_Mmo),
	(left_temp2	right_temp2) ×2,	(mouth_Ci	right_temp2),	(left_temp2	mouth_Ci),	(left_weekend1	right_weekend1) ×2,	(mouth_Mmo	right_weekend1),	(left_weekend1	mouth_Mmo),
	(left_cold1	right_cold1),	(mouth_Ci	right_cold1),	left_cold1	mouth_Ci),	(mouth_Mmo	right_date:10),	(left_day1	right_day1),	(left_until1	right_until1),
	(eye_EBf	right_danger1),	(eye_EBf	mouth_Mo1)			(left_snow1	right_snow1) ×2,	(eye_EBf	right_snow1),	(eye_EBf	left_snow1),
							(left_temp2	right_temp2) ×2,	(mouth_Ci	right_temp2),	(left_temp2	mouth_Ci),
							(left_cold1	right_cold1)				
t3	(right_tomorrow1	right_date:8	right_weather1),	(right_date:8	right_weather1	right_afternoon1),	(right_night1	right_start1	right_weekend1),	(right_start1	right_weekend1	right_date:10),
	(right_weather1	right_afternoon1	right_start1),	(right_afternoon1	right_start1	right_snow1),	(right_weekend1	right_date:10	right_day1),	(right_date:10	right_day1	right_until1),
	(right_start1	right_snow1	right_temp2),	(right_snow	right_temp2	right_cold1),	(right_day1	right_until1	right_snow1),	(right_until1	right_snow1	right_temp2),
	(right_temp2	right_cold1	right_danger1),	(left_weather1	left_afternoon	left_start1),	(right_snow1	right_temp2	right_cold1),	(right_temp2	right_cold1	right_danger1),
	(left_start1	left_snow1	left_temp2),	left_snow1	left_temp2	left_cold1)	(left_night1	left_start1	left_weekend1),	(left_start1	left_weekend1	left_day1),
							(left_weekend1	left_day1	left_until1),	(left_day1	left_until1	left_snow1),
							(left_until1	left_snow1	left_temp2),	(left_snow1	left_temp2	left_cold1),
							(mouth_Mmo	mouth_Mmo	mouth_Mmo) ×2,	(mouth_Mmo	mouth_Mmo	mouth_Ci)

Table 6: Example grams.