A New Dataset for End-to-End Sign Language Translation: The Greek Elementary School Dataset

Andreas Voskou* [1], Konstantinos P. Panousis[1], Harris Partaourides[2], Kyriakos Tolias[1], and Sotirios Chatzis[1]

[1]Cyprus University of Technology, [2]AI Cyprus Ethical Novelties Ltd

Abstract

Automatic Sign Language Translation (SLT) is a research avenue of great societal impact. End-to-End SLT facilitates the interaction of Hard-of-Hearing (HoH) with hearing people, thus improving their social life and opportunities for participation in social life. However, research within this frame of reference is still in its infancy, and current resources are particularly limited. Existing SLT methods are either of low translation ability or are trained and evaluated on datasets of restricted vocabulary and questionable real-world value. A characteristic example is Phoenix2014T benchmark dataset, which only covers weather forecasts in German Sign Language. To address this shortage of resources, we introduce a newly constructed collection of 29653 Greek Sign Language video-translation pairs which is based on the official syllabus of Greek Elementary School. Our dataset covers a wide range of subjects. We use this novel dataset to train recent state-of-the-art Transformer-based methods widely used in SLT research. Our results demonstrate the potential of our introduced dataset to advance SLT research by offering a favourable balance between usability and real-world value.

1. Introduction

Sign Language (SL) is a medium of communication that primarily uses hand gestures, facial expressions, and body movement to convey a speaker's thoughts, forming a complete and formal language. It is the primary means of communication for deaf individuals. National Sign Languages, being the native languages of deaf SL users, are a vital aspect of cultural diversity in Europe and the world. Access to SL communication is essential for HoH as it enables access to equal education, employment, and healthcare services. In Europe, there are 30 official Sign Languages and over 750,000 SL users, but only 12,000 interpreters. This shortage undermines the right to equal education and health and often endangers the lives of deaf people.

In contrast to the common misconception, Sign Languages are completely independent natural languages. Each national SL is unique with its own grammar, syntax, and vocabulary. Additionally, there is no direct connection between a spoken language and its corresponding Sign Language, for example, Greek and Greek Sign Language. A true Sign Language Translation (SLT) system needs to capture the visual patterns in the signed signal, decompose their linguistic meanings, and reconstruct them into spoken language text. These facts make the SLT task particularly challenging from a technical point of view.

Despite the importance of SLT systems, progress in this field has been limited. The current research has almost exclusively focused on simpler tasks, such as the recognition of static or dynamic gestures. Furthermore, the vast majority of successful SLT models currently available are trained on the Phoenix2014T [8], which is a single-topic dataset with a limited vocabulary, or similarly restricted datasets such as CSL-Daily [43]. While models trained on more meaningful datasets do exist, usually they either yield outcomes of significantly low quality or suffer from other types of limitations. These facts make it clear that there is a need for better machine learning techniques and, consequently, a need for more and better training datasets.

This work addresses this problem by introducing a new SLT-suitable dataset, the Greek Elementary School Dataset (Elementary23). This dataset comprises more than 28,000 videos of Greek SL, totaling over 70 hours, each paired with its corresponding spoken Greek translation in text. Prioritizing optimal technical quality, the data was captured using high-definition cameras and featured expert signers. All examples are derived from the teaching materials of Greek elementary schools, covering a broad spectrum of subjects.

*ai.voskou@edu.cut.ac.cy

2. Related Work

Datasets. Sign Language Processing involves a variety of tasks; the most common and well-studied is Sign Language Recognition (SLR) [3, 44, 6]. SLR is technically a special case of video classification, where a (short) video sequence is assigned a single label. Datasets designed for the particular task include SL videos annotated with the so-called Glosses; these are text-like labels that express discrete SL expressions. Glosses should not be mistaken for text, since they do not form a proper or complete language format and often lack expressive power. Representative SLR dataset are the DGS Kinect 40 dataset [29, 18] in German Sign Language, the GSL dataset [1] in Greek SL and many more [36, 28, 2, 15, 5, 14, 26, 40, 30, 9].

The availability of SL-related datasets is clearly considerable. However, only a select few are suitable for the most critical application of SL processing, namely end-to-end SLT. A dataset is suitable for SLT model training if it possesses two crucial characteristics: i) it includes SL videos paired with corresponding translations in a formal spoken language, and ii) the video-text pairs comprise complete and syntactically correct sentences/phrases of adequate length.

In this context, Phoenix2014T [8] dataset has become one of the most extensively studied datasets for this purpose. It features weather news performed in German Sign Language, and includes both Gloss annotation and text translations. The inclusion of Gloss annotation, combined with its dense single-topic vocabulary, renders the Phoenix2014T dataset more conducive to deep learning and has thus attracted significant attention.

Other notable datasets in this domain include SWISSTXT-NEWS [12] and VRT-NEWS [12], which cover a broader range of topics. These datasets are composed of TV-news data in German and Flemish Spoken and Sign Languages. However, the results of any end-to-end SLT attempts on these datasets have been disappointingly low, as they have failed to achieve even remotely acceptable translation quality. A recent important addition to this field was a project published by the BBC [4], consisting of a particularly large number of phrases in British Sign Language and English spoken language. The potential of this new dataset is especially high, mainly due to its extensive size. Nonetheless, end-to-end SLT results of deep learning models trained on this dataset have yet to appear in the related literature. Further examples are the CSL-Daily on Chinese SL with properties similar to Phoenix2014T and the American-SL dataset How2Sign [17] with good size and quality but lower reported results. A very recent and important addition is the OpenASL dataset [37], published in late 2022, covering 300h of American SL videos collected from online videos and demonstrating respectable results.

Sign Language Translation. Given the societal importance SLT, the field can be considered clearly underexplored. However, in recent years, a noteworthy increase of effort has taken place. The 2018 paper [8] has been the seminal work on the Phoenix2014T dataset. It implemented recurrent seq-to-seq architectures for modelling, which resulted in promising BLEU-4 scores; these ranged from 10 up to 19 for different SLT variants. In 2020, the authors of [11] utilized the power of Transformer networks in the form of the Sign Language Transformer and achieved major improvements in the translations. The approach used an S2T architecture and a feature-extracting element pre-trained as part of an SLR engine. Using the mentioned setups and additional Gloss-level supervision, they achieved clearly superior results. The 2021 paper [39] proposed a breakthrough variant of Transformers, which uses a novel form of activation functions which yield *sparse* and *stochastic* representations. This is achieved via a local stochastic competition mechanism that gives rise to stochastic local winner-takes-all (LWTA) units. The method managed to improve the translation quality even more without auxiliary Gloss supervision. In addition, they showed that the proposed method can be tuned to reduce the post-training memory footprint by properly exploiting model uncertainty. Other similar works include [10], based on a multistream Transformer, and approaches like [22] and [41] that leverage supplemental data through transfer learning and augmentation techniques to gain some additional improvement.

3. The Elementary23 Dataset

3.1. Core Elements

The introduced Greek Elementary School dataset [1], dubbed Elementary23, constitutes a noteworthy contribution to the existing body of literature due to its exceptional quality and the substantial number of examples included. Table 1 presents a comparative analysis of Elementary23 and the widely utilized Phoenix2014T dataset. As we show, Elementary23 exhibits superior technical and linguistic characteristics.

Table 1. Elementary23 statistics vs Phoenix2014T.

	Phoenix2014T	Elementary23
Signers	9	9
Total Hours	25	71
Total Frames	≈ 1.1M	$\approx 6,3$M
Sentences	8257	29653
Vocabulary	2887	23204
Singletons	1077	10126
Resolution	210×260	1280×720
FPS	25	25

[1] https://zenodo.org/record/7847052

3.2. Collection Procedure

The Elementary23 dataset stands in contrast to many relevant datasets as it was not built by annotating preexisting sign language (SL) videos from online or other sources. Instead, it was assembled through the recording of sign language interpretations of authentic elementary school content. Furthermore the final content was rigorously curated and selected by expert staff, ensuring its high impact and practical value to the deaf community and students.

The recordings for the Elementary23 dataset were made in an environment optimally suited for the task. This included a fixed single-color background and ideal lighting conditions (see Fig. 1). Professional-grade cameras and equipment were used to record videos at 720p resolution and a frame rate of 25fps.

Figure 1. Example of Elementary23 Sign Language Video Frames

The material was organised into sentences/phrases and then assigned to nine signers, all proficient users of Greek Sign Language with extensive knowledge and experience. As a result, the dataset is of exceptional quality, with minimum kinesthesiological and technical errors. The distribution per signer is illustrated in Figure 2.

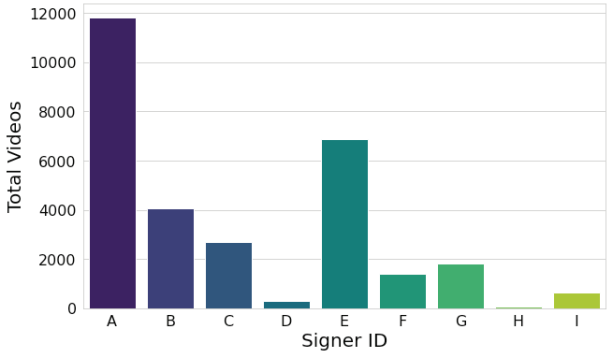

Figure 2. Distribution of video-translation pairs per signer

3.3. Content and Vocabulary

A Notable aspect of Elementary23 is its broad thematic spectrum. As previously mentioned the dataset is based on the official syllabus of Greek elementary schools, including the subjects of Greek Language, Mathematics, Religion Study, Environmental Study, History, and Anthology. The combination of those subjects ensures a sizeable lexicon of 23,204 words; yet, it is important to note that each individual subject contributes broad content and an extensive vocabulary. The statistics for each subject are presented in Table 2. The largest in terms of video-translation pairs quantity is the subject of the "Greek Language", which contains 9499 examples; the smallest one is "Religion Study", with 1,825 entries. Vocabulary-wise, the most comprehensive subject is Anthology which includes a total of 14,741 different words; on the other end, "Mathematics" have the smallest vocabulary with 6,457 words.

Table 2. Number of examples per Subject and vocabulary metrics

	Vocabulary	Examples
Anthology	14741	4158
Greek Language	14345	9499
Mathematics	6457	6583
History	7716	2067
Envir. Study	9489	5521
Relig. Study	8087	1825

3.4. SLT Subset

The main motivation of Elementary23 is to contribute one of the largest Sign Language datasets paying particular emphasis to technical and linguistic excellence. However, the dataset is not necessarily an ideal candidate for training end-to-end SLT deep networks. Specifically, there are aspects of Elementary23 that present significant modelling challenges for deep networks. These include the dataset's high word sparsity; the high number of singletons (words appearing only once in the corpus); the inclusion of particularly small phrases; and the limited number of frequently-appearing words. To be fair, this is not a problem with the dataset itself but rather a limitation of modern deep networks, which typically require multiple examples to learn from data. To overcome this issue, in this work we also present an appropriate representative subsample of the dataset, which we dub Elementary23-SLT. This is more suitable for training end-to-end SLT deep networks, and has a size similar to Phoenix2014T and other recently published benchmark datasets.

The selection process was guided by three primary principles: (i) decrease the absolute number of singletons; (ii) increase the density of frequent words and bigrams; and (iii) keep content diverse. To achieve these targets, we first went through preliminary cleaning, whereby we eliminated singleton-only sentences. Afterwards, we ran a simple dynamic multi-round elimination process: At each round, we listed all sentences containing singletons but no frequent el-

Table 3. Elementary23-SLT vs key Benchmarks

	Elementary23-SLT	Phoenix2014T [8]	SWISSTXT-NEWS [12]	VRT-NEWS [12]
Sentences	8372	8257	6031	7174
Total Words	83327	99081	72892	79833
Vocabulary	8202	2887	10561	6875
Mean Word Freq.	10.16	34.3	6.9	11.6
Sigletons	3327 (41%)	1077 (37%)	5969 (57%)	3405 (50%)
Rare words (<5)	6155 (75%)	1758 (60%)	8779 (83%)	5334 (78%)

ements, and we eliminated approximately a quarter of them. We then recalculated word frequencies and redefined singletons and frequent words based on the surviving subset. At the end of each round, we confirmed that all six subjects were represented with a sufficient number of remaining examples, at least 10% of the original. We repeated this procedure several times, until we ended up with a sample size comparable to the standard benchmarks.

The aforementioned process left us with a sample of 7168 sentences. Subsequently, we split the data into the typical train, validation and test subsets. During the selection of the validation and test sets, we tried to avoid sentences shorter than four words long, and made sure to exclude all the sentences appearing twice by the same speaker. As an extra processing step, we augmented the training set with an additional 1,204 non-singleton single-word videos. Finally, for comparison reasons, we additionally performed a train-validation-test split on the entire dataset following similar principles regarding duplicated entries. We will be referring to this split as Elementary23-Raw. In both cases, all speakers may participate in all the subsets.

The final SLT set contains 8372 video-sentence pairs, organized as 7348 pairs for the training set, 512 for the validation set, and 512 for the test set. Through this operation, we produced an effective subset in the typical size spectrum, that retains the desired qualitative elements of the complete collection while exhibiting some improved quantitative factors. Key statistics regarding the subset and benchmarks are stated in Table 3 and will be analysed later in this Section.

To enable the examination and exploitation of the data by the research community, we have ensured that Elementary23-SLT complies with established standards regarding size and structure. Additionally, the data will be available in a file format that is practical and consistent with prior works [11, 39]. Specifically, we have created a JSON file comprising a list of dictionaries, each corresponding to a particular video-sentence pair. These dictionaries encompass all auxiliary elements, such as numbering and signer ID, as well as the principal input-output data; the latter consists of the frame-wise feature sequence and the corresponding translation in modern Greek.

While feature extraction is technically a component of deep network development, we have adopted conventional practice by embedding the extracted features within the introduced SLT dataset. This approach does not affect the deep network training process or the final product, yet it considerably reduces the costs and effort of model development. The complete videos will also be made available.

3.5. Landmarks and Trajectories

State-of-the-art SLT networks often utilize convolutional subparts as feature extractors; these are pre-trained on sign language recognition datasets. Such subparts can extract spatial information from video frames by leveraging their prior knowledge of core sign language elements. However, this preprocessing phase requires laborious dataset annotation in terms of auxiliary Glosses. Therefore, this process is of reduced applicability: developed models can only generalize on other datasets of similar Gloss structure. In addition, it is incompatible with the Greek Elementary School dataset: its immense size renders provision of Glosses completely out of scope.

To address this issue, we follow an alternative, yet established approach that involves using the OpenPose engine [13]. The OpenPose engine is a convolutional neural network that has been trained to track and extract the trajectories of key human body parts. We use OpenPose to track landmarks related to the 2D positioning of upper body movements, facial expressions, and hand shapes; one can use these as input to a developed end-to-end SLT model. We further scrutinize the extracted features, by excluding components that appear not to contribute enough or exhibit persistently low volatility, such as lower body landmarks. The resulting vector effectively summarizes the video frames and serves as a sufficient source of information for subsequent network layers. Figure 3 illustrates a representative example of this approach.

3.6. Lexical Statistics and Benchmarks

While Phoenix2014T has gained popularity as a standard benchmark for SLT due to its ease of modelling, its appropriateness as a benchmark for comparing to the newly introduced dataset remains questionable. This is due to the following facts: i) it covers only a single topic, in contrast to the multi-subject nature of Elementary23; ii) vocabulary coverage is very limited; iii) it includes many sentences of

Figure 3. Body Landmarks - Trajectories

similar structure and content, due to the weather forecasting's strict format; and iv) Phoenix2014T includes auxiliary Gloss annotation.

To address these challenges, we're supplementing our benchmark schema with two more datasets: SWISSTXT-NEWS and VRT-NEWS. Containing 6031 and 7174 videos/sentences respectively, they're based on HD TV news videos. We chose them due to their similar size to Phoenix2014T and Elementary23-SLT, good video quality and their basis in European sign languages, ensuring directly comparable grammatical structures.

In Table 3, we present a detailed comparison of Elementary23-SLT and the three selected benchmark datasets on the grounds of various vocabulary-oriented metrics. These metrics were critical in determining the suitability of SWISSTXT-NEWS and VRT-NEWS as the primary benchmarks, since their statistics closely resemble those of Elementary23-SLT. Specifically, both SWISSTXT-NEWS and VRT-NEWS contain similarly sized vocabularies, with 10561 and 6875 sentences, respectively. Thus, they deviate no more than 25 % from our proposed subset of 8202 sentences. Furthermore, the percentage of rare words, defined as words that appear less than five times, ranges from 75 % to 83% for all three datasets. Finally, the mean word frequencies are comparable across these datasets, with the words in SWISSTXT-NEWS appearing an average of 11.6 times, 6.9 for VRT-NEWS, and 10.1 for Elementary23-SLT.

Notably, Phoenix2014T has a vastly reduced vocabulary size compared to the rest of the considered datasets, which constitutes a central constraint. Additionally, Phoenix2014T is characterized by a considerably higher mean word frequency equal to 34.3. This number is rooted in the limited vocabulary and allows for easier training since it narrows down verbal varieties. Furthermore, Phoenix2014T has the lowest percentage of rare words and singletons. In terms of singletons, Elementary23-SLT is a close second, with an increase of only 4%.

4. Translation Methodology

In order to tackle the challenging task of end-to-end SLT, we adhere to the guidance of recent developments in the field that advocate for the use of Transformer-based architectures [11, 39, 42]. Specifically, we employ the seminal SLT model of [11], and the later sLWTA-Transformer [39] variant; we focus more on the latter method, due to its many advantages. These techniques have been demonstrated to be effective on the well-known PHOENIX-14T dataset, and achieve state-of-the-art results with BLEU-4 scores in the area of 22 and 24, respectively.

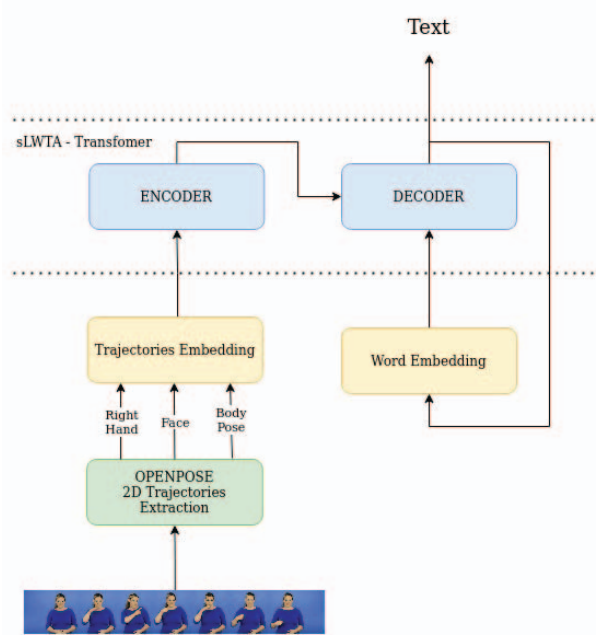

Figure 4. The suggested Sign to Text Transformer Network

Transformers are modern deep architectures that rely purely on the Attention mechanism to process temporal dynamics in sequential observations. A standard Transformer layer includes a self-attention layer, paired with an immediately succeeding Relu-activated Feed-Forward Network (FFN). Sign language Transformers comprise encoder and decoder parts similar to [38]. The encoder is presented with frame-wise feature vectors obtained from a spatial feature extractor; it learns to process their interactions over the temporal axis and yields action-aware representations. The decoder uses the latter and re-expresses the meaning into formal spoken language.

The sLWTA SL-Tranformer reapproaches the standard architecture by introducing two forms of stochasticity: (i) Gaussian weights with posterior distributions estimated through variational inference, instead of standard point estimators on the weights; and (ii) the stochastic local-winner-takes-all layer as a more sophisticated FFN.

In more detail, we consider a Bayesian treatment of the

weights by training a Gaussian variational posterior; this encloses an uncertainty estimation of each weight, formed as $q(w) = N(\mu, \sigma^2)$ where μ, σ^2 are the posterior mean and variance. For inference, weight values are sampled from the Gaussian posteriors in a Monte Carlo fashion. The reparameterization trick of [24] is employed to allow for gradient descent-based training.

The stochastic local-winner-takes-all layer [31], reported as LWTA or sLWTA, is a sophisticated non-linear layer that replaces the usual relu-activated dense layers with notable success [34, 32, 23, 33]. Let $\boldsymbol{x} \in R^J$ be the input of a typical dense layer, and $\boldsymbol{y} \in R^H$ the corresponding output vector gained through multiplication with a weight matrix $W \in R^{J \times H}$ and activation via a nonlinear function such as ReLU $\boldsymbol{y} = ReLU(W\boldsymbol{x})$. LWTA works by organising the output $\boldsymbol{y}$ into K blocks of U members/competitors each, and the weight matrix W into K respective submatrices. For any block indicated by $k \in \{1, 2, ..K\}$, we denote as $\boldsymbol{y}_k \in R^U$ and $W_k \in R^{J \times U}$ the corresponding subparts of y and W. The members of each block compete with each other, and only one of them, the winner, gets activated given an input, while the rest are set to 0. The so-called competition is an inter-block stochastic procedure based on sampling the winner from a discrete posterior with logits proportional to the linear computation in each unit:

$$q(\boldsymbol{\xi}_k) = \mathrm{D}\Big(\boldsymbol{\xi}_k \Big| \mathrm{softmax}\big(W_k \boldsymbol{x}\big)\Big), \quad \forall k \in \{1, 2, ..K\} \quad (1)$$

where $\boldsymbol{\xi}_k \in onehot(U)$ are discrete latent one-hot vectors indicating the winner of each block.

The sampling process is effectively approximated using the Gumbel-Softmax relaxation trick [21] . Controlled by a temperature hyper-parameter T, this technique can offer low-variance gradients during the training phase (high T), and hard discrete samples, almost identical to Eq. (1), during inference (low T).

Finaly, using the postulated $\boldsymbol{\xi}$ latent variables, layer output y can be expressed as follows:

$$\boldsymbol{y}_k = \boldsymbol{\xi}_k \odot (W_k \boldsymbol{x}_k), \quad \forall k \in \{1, 2, ..K\} \quad (2)$$

where $\odot$ stands for element-wise multiplication.

4.1. Training and Inference

In the following, we will be referring to the standard Transformer-based SLT model of [11] as the deterministic model. On the other hand, the variant of [39] will be dubbed as the stochastic model.

The training objective of the deterministic SLT model will be to minimise the cross-entropy error between each predicted word and the corresponding label, under a standard seq-to-seq rationale. When it comes to inference using the trained model, we again follow the usual practice and run autoregressive decoding, where words of each produced sentence are predicted one by one, and the decoder is presented the encoded representations and the previous predictions. Additionally, we use the beam search algorithm, the parameters of which are optimised on the validation set.

Training of the stochastic variant is slightly more complex. The optimization objective is the negative evidence lower bound (ELBO) of the model, computation of which requires prior assumptions regarding the distributions of the winner indicator latent variables, $\boldsymbol{\xi}$ on each LWTA layer, as well as the trainable weights, $\boldsymbol{w}$, throughout the network. For convenience, we postulate a priori spherical Gaussian weights of the form $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and a symmetric Discrete prior over the winners: $p(\boldsymbol{\xi}) = \mathrm{Discrete}(1/U)$.

Then, the training objective comprises: (i) the standard cross-entropy of the network, with the expressions of the latent variables, $\boldsymbol{\xi}$, expressed via the Gumbel-softmax reparameterization trick, and the Gaussian weights, $\boldsymbol{w} \sim q(\boldsymbol{w})$, expressed via the standard reparameterization trick for Gaussians; (ii) the Kullback-Leibler divergences between the posterior and the prior of the latent variables, $\boldsymbol{\xi}$, and the Gaussian weights, $\boldsymbol{w}$ [39].

For inference, we directly draw weight samples $\boldsymbol{w}$ from the trained Gaussian posteriors. Similarly, we directly draw samples of the winner-indicating latent vectors, $\boldsymbol{\xi}$, from the related discrete posteriors. The final prediction is obtained through Bayesian averaging; we sample the weights and $\boldsymbol{\xi}$ parameters from the posteriors four times, calculate the final logits, and average the results. This way we obtain the final logits that we use to drive beam search, similar to the deterministic model.

5. Experimental Results

5.1. Experimental setup

This section presents our experimental results, primarily focusing on the LWTA-Transformer's application on the SLT subset of the Elementary dataset. The suggested LWTA-Transformer version is a three-layer architecture with an embedding size of 256 and U=2 competing units.

Following the recommendations of [11], and [35] we initialised the trainable parameters using Kaiming uniform for stochastic models [20], and Xavier normal for deterministic models [19]. The Gumbel-Softmax temperature was set following the related theory in [21]; we use a high temperature of $T = 1.00$ during training and a low $T = 0.01$ during inference. The rest of the training hyperparameters were either chosen based on the exact suggestions from the original papers of the models or optimised during our experimental investigation. The BLEU-4 score was used as the main evaluation metric to assess the quality of sign-to-text translation. Core parts of the model's implementation are modified versions of [25, 11, 39].

Table 4. Results using deterministic and stochastic Transformers for both the entire dataset and SLT subset

Data	Model	Dev				Test			
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Raw	Stochastic	10.4	2.14	0.95	0.33	11.50	2.85	1.05	0
	Deterministic	6.23	1.23	0.54	0.36	7.68	1.83	0.5	0
SLT	Stochastic	21.30	12.26	8.74	**6.67**	19.99	11.10	7.68	5.69
	Deterministic	18.79	9.69	6.68	5.08	17.37	8.50	5.35	3.85

5.2. Quantitative Results

Two central inquiries targeted in our experiments are (i) the identification of the network architecture that produces the best outcomes; and (ii) quantifying the impact of our decision to use an SLT-suitable subset rather than working with the entire Elementary23 dataset. Table 4 presents the best results that were achieved for all the subcases.

Table 4 clearly demonstrates that using the Raw dataset resulted in very low BLEU scores for both the deterministic and stochastic models. Conversely, the usage of the SLT-subset resulted in significantly improved outcomes. The exclusion of problematic or unsuitable sentences and an appropriate data split apparently play a crucial role, making training less noisy and more focused on effective examples.

In terms of architecture, we compare two of the best approaches available, that is, the original S2T deterministic Transformer and its stochastic counterpart, i.e. the sLWTA Transformer. The results are summarised in Table 4. The stochastic LWTA Transformer appears to be superior to the deterministic, achieving a BLEU-4 score of more than 1.5 units higher than the latter. The superiority holds for all ranks of BLEU scores in both the training and validation set.

5.2.1 Ablation study

Model size Contemporary NLP Transformer networks have a tendency towards large size [27, 7, 16], reaching depths that can approach 100 layers. Conversely, SL Transformers are commonly much smaller, often no more than 2 to 3 layers deep. Our experiments, conducted on the Elementary23-SLT dataset, aimed to determine the optimal depth for the proposed stochastic Transformer. The results, depicted in Table 5, demonstrate that a depth of 2 proves to be the optimal choice, as indicated by the highest BLEU-4 scores on both the test and validation(dev) sets. A depth of 1 delivered results that were lower but still close, while increasing the depth beyond 2 resulted in a decline in performance of around 1 to 1.5 units. The embedding size is another size-related hyperparameter crucial in Deep NLP models. Table 6 presents a study of the effect of different embedding sizes on the performance of the LWTA-Transformer on the Elementary23 dataset. The results indicate that an embedding size of 256 provides optimal performance. Smaller sizes appear insufficient to handle the complexity of the task, as they yielding the much lower scores of 4.39/4.07 for size=128. On the other hand, larger sizes, such as 512, do not improve the results.

Table 5. BLEU-4 Scores per model depth

Depth	Dev	Test
1	6.35	5.57
2	6.67	5.69
3	5.09	4.63

Table 6. BLEU-4 Comparison between embedding sizes

Embedding Size	Dev	Test
128	4.39	4.07
256	6.67	5.69
512	5.32	5.27

The effect of Competing Units per Block As previously discussed, the results of our experiments on the sLWTA Sign Language Transformer (Table 4) render it a superior solution for the Greek SL translation task compared to the deterministic model. A central aspect of this network is the use of the sophisticated LWTA layer, as opposed to a typical activation function such as Relu. The size of the competition blocks U is the main tunable hyperparameter of this technique. Through an examination of the commonly used sizes, presented in Table 7, we concluded that the most suitable choice for our case is $U = 2$, as suggested in [31]. Larger sizes of $U = 4$ and $U = 8$ resulted in decreases of 0.22 and 0.76 BLEU-4 units, respectively; this is likely due to the high sparsity of the representations obtained from the LWTA blocks.

Table 7. The effect of LWTA block size U

Competing Units (U)	Dev	Test
2	6.67	5.69
4	5.41	5.47
8	5.23	4.93

5.2.2 Discussion and Benchmarking

We now proceed with a direct evaluation of the translation accuracy attained on the Elementary23 dataset, and compare it to the results reported on other benchmark datasets. Table 8 summarises the achieved BLEU-4 scores for each case, covering both the main benchmarks and a curated selection of significant yet less directly comparable supplementary non-European datasets.

Table 8. Benchmarking BLEU-4 scores

Dataset	Dev	Test
Elementary23-SLT	6.67	5.69
Phoenix2014T [39]	23.23	23.65
SWISSTXT-NEWS [12]	0.46	0.41
VRT-NEWS [12]	0.45	0.36
OpenASL [37]	6.57	6.72
CSL-Daily [43]	20.80	21.34

As indicated in this table, researchers have reported BLEU-4 scores reaching as high as 23.23/23.65 for the validation and test sets of Phoenix2014T, validating its standing as the highest-performing dataset. However, as previously noted, this dataset does come with limitations such as a restricted vocabulary, a narrowly focused topic, and a stringent structure. Analogous results emerge from applications on the CSL-Daily dataset. More specifically, researchers report impressive scores of 20.80/21.34 on this popular Chinese-SL dataset, which also bears similar constraints on vocabulary and content to Phoenix2014T. While these attributes enhance BLEU-4 performance, they diminish the applicability of the developed SLT models for real-world users. In contrast, the objective of our work is to amplify the effectiveness of end-to-end SLT systems in genuine usage scenarios. These considerations make Phoenix2014T an imperfect comparison to the Elementary23 dataset, which boasts a more realistic design.

Conversely, models trained on SWISSTXT-NEWS and VRT-NEWS exhibit weak performance, with BLEU-4 scores < 1. The authors report scores of 0.46/0.41 and 0.45/0.36, respectively, which are considerably lower than the 6.67/5.69 achieved in our proposed subset. Unfortunately, with such poor scores, these datasets cannot provide any practical value, nor can they be regarded as a reliable benchmark for future SLT models; these facts are despite their extensive topic coverage and vast vocabulary size.

These contrasting outcomes underscore the value of the new Elementary23 dataset. Although the attained BLEU-4 scores, around 6 units, are low compared to state-of-the-art NLP models trained on extensive text corpora, they demonstrate tangible translation capabilities. Therefore, unlike the previously mentioned SL datasets, our data combine measurable results with a comprehensive and realistic thematic spectrum. These qualities give our proposed dataset particular importance as a benchmark dataset for future SLT research. Finally, the more modern OpenASL is the only case that seems to align with Elementary23, combining respectable results with high-quality content.

5.3. Qualitative Results

The quality of the automatic translations produced by our models varies; this is shown in Table 9, where three representative examples are presented. In some cases, such as the first example, the results are impressively accurate, with only minor numerical or grammatical errors. The next case belongs to a second category in which the context is partially captured, but the syntax deviates from the target. The final example represents the third group, where the model completely fails to detect any of the signed signals.

Table 9. Reference (R), Prediction (P), Translated Reference (Rt), Translated Prediction (Pt)

#	Translation Reference and Prediction
1	**R:** έχει 10 νομίσματα πόσα είναι τα χρήματά του συνολικά **P:** έχει 4 νομίσματα πόσα είναι τα χρήματά του συνολικά **Rt:** he has 10 coins how much is his money in total **Pt:** he has 4 coins how much is his money in total
2	**R:** συμπληρώνω τους αριθμούς που λείπουν στους πίνακες **P:** υπολογίζω και γράφω τους αριθμούς που λείπουν **Rt:** I fill in the missing numbers in the tables **Pt:** I calculate and write the missing numbers
3	**R:** στο σχολείο μαθαίνουμε καινούρια πράγματα **P:** το περιβάλλον μου **Rt:** at school we learn new things **Pt:** my environment

6. Conclusions

This paper presents a novel dataset of Greek Sign Language, characterized by high technical quality and diverse vocabulary. Using a simple yet effective selection procedure, we selected an SLT-suitable subset that adheres to established formatting standards and retains the desired features. By utilizing variants of the state-of-the-art LWTA Transformer, we achieved BLEU-4 scores ten times higher than those reported on directly comparable datasets.

As such, our proposed Elementary23 dataset offers a balance of quality and feasibility. Elementary23 can serve as a viable alternative to popular yet monotonous dataset like Phoenix2014T and similar, and the verbally rich but unattainable such as SWISSTXT-NEWS and VRT-NEWS options. Future work could involve applying even more sophisticated SLT models, exploring data-oriented techniques such as active or meta-learning, and further investigating the use of the Elementary23 dataset for tasks such as text-to-sign SL production.

References

[1] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2(2), 2020.

[2] Ulrich von Agris and Karl-Friedrich Kraiss. SIGNUM database: Video corpus for signer-independent continuous sign language recognition. In Philippe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martínez Ruiz, and Adam Schembri, editors, *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 243–246, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[3] Mohammed S Al-Samarraay, Mahmood M Salih, Mohamed A Ahmed, AA Zaidan, Osamah Shihab Albahri, Dragan Pamucar, HA AlSattar, Abdullah Hussein Alamoodi, BB Zaidan, Kareem Dawood, et al. A new extension of fdosm based on pythagorean fuzzy environment for evaluating and benchmarking sign language recognition systems. *Neural Computing and Applications*, pages 1–19, 2022.

[4] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

[5] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[6] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 182–191, 2022.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[9] Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. BosphorusSign: a Turkish sign language recognition corpus in health and finance domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1383–1388, 2016.

[10] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 301–319. Springer, 2020.

[11] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. IEEE, 2020.

[12] Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[14] X Chai, H Wanga, M Zhoub, G Wub, H Lic, and X Chena. Devisign: dataset and evaluation for 3d sign language recognition. *Technical report, Beijing, Tech. Rep*, 2015.

[15] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. *Sign Language Recognition Using Sub-units*, pages 89–118. Springer International Publishing, Cham, 2017.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744, 2021.

[18] Ralph Elliott, Helen Cooper, John Glauert, Richard Bowden, and François Lefebvre-Albaret. Search-by-example in multilingual sign language databases. In *Proceedings of the Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, Scotland, Oct. 23 2011.

[19] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

[22] Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. Prior knowledge and memory enriched transformer for sign language translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775, 2022.

[23] Konstantinos Kalais and Sotirios Chatzis. Stochastic deep networks with linear competing units for model-agnostic meta-learning. In *International Conference on Machine Learning*, pages 10586–10597. PMLR, 2022.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[26] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1975–1979, 2012.

[27] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[28] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[29] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sequential pattern trees. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, June 16 – 21 2012.

[30] Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgoz, and Lale Akarun. BosphorusSign22k Sign Language Recognition Dataset. 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*, 2020.

[31] Konstantinos Panousis, Sotirios Chatzis, and Sergios Theodoridis. Nonparametric bayesian deep networks with local competition. In *International Conference on Machine Learning*, pages 4980–4988. PMLR, 2019.

[32] Konstantinos Panousis, Sotirios Chatzis, and Sergios Theodoridis. Stochastic local winner-takes-all networks enable profound adversarial robustness. In *Bayesian Deep Learning NeurIPS workshop*, 2021.

[33] Konstantinos P Panousis, Anastasios Antoniadis, and Sotirios Chatzis. Competing mutual information constraints with stochastic competition-based activations for learning diversified representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7931–7940, 2022.

[34] Konstantinos P. Panousis, Sotirios Chatzis, Antonios Alexos, and Sergios Theodoridis. Local competition and stochasticity for adversarial robustness in deep learning. In *Proc. AISTATS*, 2021.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[36] Franco Ronchetti, Facundo Quiroga, Cesar Estrebou, Laura Lanzarini, and Alejandro Rosete. Lsa64: A dataset of argentinian sign language. *XX II Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016.

[37] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*, 2022.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[39] Andreas Voskou, Konstantinos P. Panousis, Dimitrios Kosmopoulos, Dimitris N. Metaxas, and Sotirios Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proc. ICCV*, 2021.

[40] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 872–885, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[41] Pan Xie, Mengyi Zhao, and Xiaohui Hu. Pisltrc: Position-informed sign language transformer with content-aware convolution. *IEEE Transactions on Multimedia*, 24:3908–3919, 2021.

[42] Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.

[43] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021.

[44] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5131–5140, 2022.