

Learning Sign-Oriented Intermediate Reordering from Dependency-Based Weak Supervision

by

Daniyal Bayanov

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

June 2026

© Nazarbayev University 2026. All rights reserved.

Author
Department of Computer Science
June 2026

Certified by
Anara Sandygulova
Associate Professor
Thesis Supervisor

Accepted by
Adnan Yazici
Chair, Department of Computer Science

Learning Sign-Oriented Intermediate Reordering from Dependency-Based Weak Supervision

by

Daniyal Bayanov

Submitted to the Department of Computer Science
on June 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

The challenge in translating speech into sign language is not only the different vocabulary used, but also the different structure of both languages. It becomes particularly hard in low-resource settings when no large parallel corpora are available and glossing, notation, or even aligned outputs of the sign language are scarce. Instead, direct one-to-one word translation results in unnatural intermediate forms in sign generation tasks. In this thesis, we address a related but smaller problem of predicting the order of signs in sign-oriented intermediates generated by a deterministic dependency-based rule system. Our goal is not to translate speech into sign language, but to find an appropriate order of tokens for subsequent sign generation.

We construct pseudo-targets using the Universal Dependencies annotation of the Russian SynTagRus treebank. The final dataset consists of 6,000 filtered sentences obtained from 69,631 treebank sentences via several filtering steps. These include removal of sentences with extreme length, questions, subordinate clauses, and complex multi-clause sentences. The final split includes 4,800 training, 600 validation, and 600 test sentences.

Two multilingual sequence-to-sequence transformers are trained to mimic the output of the rule system. The baseline, mT5-small, obtains a test set BLEU score of 52.57 ± 0.49 and chrF score of 80.86 ± 0.44 , and exact match of 0.27, token accuracy of 0.52, and Kendall’s τ of 0.80. A larger variant, mT5-base, boosts the performance to BLEU of 63.90, chrF of 87.53, exact match of 0.35, token accuracy of 0.62, and Kendall’s τ of 0.86. The improvement by 11 BLEU points proves that the model capacity plays a role in the task and there is room left for further improvements. Both models manage to capture the structural characteristics defined by the rule system, namely the preference of placing the subject before the verb and the object before the verb.

The thesis also provides a multi-seed analysis of our models, a quantitative error analysis, and several ablative analyses. The error analysis demonstrates that both the number of repeated words and omissions significantly increases with sentence length, growing from 14.4% on short sentences to 88.0% on long sentences. From

the ablations, we conclude that punctuation must be preserved, while auxiliaries and adverbials do not have to be fronted in all cases. The target format is not crucial, and the dataset quality plays a more significant role than most rule choices. The contributions of this thesis lie in building a reproducible and interpretable module, which can then be included into the text-to-sign pipeline to assess the intermediate performance before sign generation.

Thesis Supervisor: Anara Sandygulova

Title: Associate Professor

Acknowledgments

Contents

1	Introduction	15
1.1	Background and Motivation	15
1.2	Problem Statement	16
1.3	What Is an Intermediate Representation?	17
1.4	Why This Is Not yet Sign-Language Generation	17
1.5	Weak Supervision and Its Role	18
1.6	Research Questions	18
1.7	Contributions	18
1.8	Thesis Structure	19
2	Related Work	21
2.1	Overview	21
2.2	Sign-Language Translation as a Distinct Structural Problem	22
2.3	Intermediate Representations in Sign-Facing Systems	23
2.4	Weak Supervision and Data Scarcity in Low-Resource Settings	24
2.5	Syntax-Aware Reordering and Structural Transfer	25
2.6	Syntactic Resources: Universal Dependencies	26
2.7	Relation to Earlier Pipeline Work	26
3	Methodology	29
3.1	Introduction	29
3.2	Notations and Terminology	30
3.3	Task Definition	30

3.4	Dataset Construction and Filtering	31
3.4.1	Syntactic Backbone	31
3.4.2	Quantitative Filtering Pipeline	31
3.5	Weakly Supervised Target Induction	34
3.5.1	Motivation	34
3.5.2	Rule Design	34
3.5.3	Linguistic Rationale and Scope of the Rules	36
3.5.4	Examples	37
3.6	Neural Reordering Models	38
3.6.1	Why a Neural Model Is Needed	38
3.6.2	Model Family and Capacity Comparison	39
3.6.3	Training Configuration	40
3.6.4	Why These Metrics Are Appropriate	40
3.6.5	Reproducibility Considerations	42
4	Experiments and Results	43
4.1	Experiments Overview	43
4.2	Baselines and Evaluation Protocol	43
4.3	Main Test-Set Results	44
4.4	Effect of Model Capacity	45
4.5	Sensitivity to Random Seed Initialization	46
4.6	Sentence Length Analysis	47
4.7	Learning Dynamics	48
4.8	Error Analysis	49
4.9	Sample Examples	50
4.10	Ablation Study	52
5	Discussion and Conclusion	55
5.1	Summary of Findings	55
5.2	Interpretation of the Learned Models	56
5.3	What the Results Mean for the Proposed IR	57

5.4	Relation to the Earlier Pipeline	58
5.5	Limitations	58
5.6	Future Work	59
5.7	Conclusion	61

List of Figures

3-1	The basic idea behind the proposed pipeline. From a spoken-language sentence annotated with dependencies, we create a pseudo-target IR through a rule system, which serves as a label for a text-to-text machine-learning system.	29
3-2	Distribution of sentences by length after filtering.	33
3-3	Split composition by sentence length buckets.	33
4-1	mT5-small vs. mT5-base on the main test-set evaluation metrics. . .	45
4-2	Visualizations of BLEU and chrF variability across mT5-small seeds.	46
4-3	Visualizations of Kendall’s τ and exact match variability across mT5-small seeds.	46
4-4	Performance by sentence length buckets for BLEU and τ	47
4-5	Training loss and epoch-wise generation metrics for mT5-small	48
4-6	Training loss and epoch-wise generation metrics for mT5-base	48
4-7	Error rates for repetition, omission, hallucination, and length mismatch.	50
4-8	Comparative analysis of different preprocessing strategies and target representations.	53

List of Tables

3.1	Detailed filtering statistics.	32
3.2	Dependency-based rule system used to generate pseudo-targets. . . .	35
4.1	Main test-set results. <code>mT5-small</code> reports mean $\pm$ standard deviation over three seeds.	44
4.2	Performance by sentence-length bucket for <code>mT5-small</code> (best seed). . .	47
4.3	Error rates for <code>mT5-small</code> on the test set, overall and by sentence-length bucket.	49
4.4	Representative qualitative examples from the test set. Examples are selected to illustrate exact matches, omission errors, and the effect of model capacity.	51
4.5	Ablation study results, aligned with the main experiment.	52

Chapter 1

Introduction

1.1 Background and Motivation

Translation between spoken and sign languages continues to be among the most important open problems in language accessibility. For deaf and hard-of-hearing individuals, information access can hinge upon whether or not spoken or written text can be mapped into a sign-language-oriented format. While much of the engineering effort devoted to this topic tends to focus on downstream visual outputs, including videos, avatar motion, or sequences of poses, the preceding linguistic structure is often a harder problem. When the text is mapped word by word without regard for the structure of the sentence, it may retain too much of the structure characteristic of spoken-language use and thus end up being less natural in the sign language.

This thesis addresses just that sort of linguistic challenge. Specifically, it focuses on one particular aspect of a larger text-to-sign pipeline: the reordering of spoken language into sign language. Rather than attempting a full-fledged solution to sign-language generation as an end-to-end problem, this thesis considers the task of reordering as an independent learning problem. Methodologically, this allows for greater precision and evaluation. Practically, it makes the task easier to experimentally constrain, improve, and evaluate in its own right, prior to adding downstream lexicalization, notation, pose generation, or rendering as an animation.

The broader motivation is based on low-resource sign-language contexts, includ-

ing those that pertain to Kazakhstan and Kazakh-Russian sign language. In such contexts, large annotated parallel datasets are costly and challenging to acquire. Previous system-building work had demonstrated the feasibility of a multi-stage processing pipeline whereby text was reduced to some base form, words retrieved from a sign-language database, converted to pose representation, and ultimately rendered as animations or GIFs. That work also revealed the inherent limitation of lexical lookup alone, insofar as the structure of the input remained misaligned with that of sign language. From this, the present thesis argues that the insertion of an intermediate reordering step can fill this gap.

1.2 Problem Statement

The primary problem addressed in this thesis is sign-oriented intermediate reordering. Put informally, the goal is to map a spoken-language sentence to an equivalent in which the token order approximates the order expected for sign language. Formally, let

$$x = (x_1, x_2, \dots, x_n)$$

represent a spoken-language sentence in textual form, and let

$$y = (y_1, y_2, \dots, y_n)$$

represent the target reordering. The task consists of learning a function that maps a spoken-language sentence to a new sequence in which the tokens maintain their lexicographical identity, but are reorganized in accordance with sign-language standards.

This is a text-to-text structural transduction task. It is neither gloss prediction, nor SignWriting generation, nor pose or video generation. The target remains textual, but with a changed token order aimed at supporting later sign-language processing.

1.3 What Is an Intermediate Representation?

An important concept in this thesis is that of Intermediate Representation, or IR. As understood here, an IR refers to any sort of representation that falls between the raw spoken-language text input and some sign-language-oriented downstream format. The important qualification here is that the IR does not refer to the final product, the sign language proper, but a kind of intermediate format designed to make subsequent processing easier. In the present thesis, this intermediate representation takes the form of a reordered sequence of tokens. What is important about this design choice is not that this sequence is necessarily sign language, but that it serves better than the original spoken language as a starting point for other components.

The advantage of this design choice is several-fold. First, it simplifies the task and makes it easier to evaluate than complete sign-language generation. Second, it allows the use of dependency-annotated corpora for training, regardless of the lack of a sign-language corpus. Finally, it enables modular system building.

1.4 Why This Is Not yet Sign-Language Generation

One of the most critical distinctions made in this thesis regards its scope. The methodology developed here does not involve direct mapping from spoken language to gloss, to SignWriting, to pose representation, to animation, or any number of similar outputs. Instead, it involves solving a smaller, isolated problem: reordering spoken language to a sign-language-like intermediate form. What matters about this distinction is that it prevents the claims made in this thesis from exceeding the actual experimentation carried out.

Thus, this thesis will not claim gold-standard accuracy, nor perfect grammatical sign structure, nor avatar quality improvement or validation, nor any native signer tests regarding end-user comprehension. Instead, what is claimed is that a sign-language reordering model can be learned through weak supervision, and that this module can greatly outperform the use of unaltered spoken-language text.

1.5 Weak Supervision and Its Role

A second important concept in this thesis is weak supervision. While supervised learning generally refers to a situation wherein the model receives training examples, along with the correct human annotation for each, there are many situations where such human annotations cannot be obtained in sufficient quantity. For example, parallel datasets annotated for a sign language in a low-resource context would be very challenging to acquire.

This thesis uses weak supervision, whereby the targets are generated through a heuristic rule system. The auto-generated targets are referred to as pseudo-targets. The task is still supervised in the machine-learning sense, in that the model is training against paired inputs/outputs, but it lacks the strong element of gold standard supervision since the target is weakly generated rather than annotated.

1.6 Research Questions

The three research questions guiding this thesis are as follows. Can a model learn the approximate sign-oriented reordering function using a dependency-based weak supervision system? Does such a model preserve the structure embedded in the rule system, rather than simply producing similar-looking strings? And finally, does capacity matter in the performance achieved?

The first two questions concern learnability and structural fidelity, respectively. While the former was known to be feasible, the latter was the main discovery made through experimentation. The third question arose as an important new result whenever evaluating this task in relation to different model sizes, since the large performance difference between mT5-small and mT5-base indicated the lack of saturation.

1.7 Contributions

The contributions of this thesis are summarized as follows.

1. Definition of sign-oriented intermediate reordering as a separate task.
2. Introduction of a dependency-based weak-supervision system for pseudo-target generation, based on Universal Dependencies annotation.
3. Demonstration of the ability of a neural network to solve this mapping function and outperform the baseline by a substantial margin.
4. Introduction of the notion of model capacity as an important factor for performance, as evidenced by the significant difference between mT5-small and mT5-base.
5. Multi-seed analysis of variance and a quantitative analysis of error patterns, including repetition, omission, hallucination, and length mismatch.
6. An ablation study, showing which properties of the target space matter for performance, particularly cleanliness of data and punctuation.
7. Situating the learned reordering model in a larger text-to-sign translation pipeline, inspired by previous engineering work.

1.8 Thesis Structure

The rest of this thesis is organized as follows. Chapter 2 reviews the literature regarding sign language translation, intermediate representations, weak supervision, and syntax-sensitive reorderings. Chapter 3 introduces the problem formulation, notation used, data construction pipeline, filter pipeline, pseudo-target generation, and neural network modeling. Chapter 4 outlines the experiments performed, including main results for both mT5-small and mT5-base, seed variation analysis, error analysis, qualitative behavior, and alignment-based ablation study. Chapter 5 discusses the interpretation of the results, as well as limitations, directions for future work, and additional implications.

Chapter 2

Related Work

2.1 Overview

The present chapter locates the proposed research in relation to several research streams: sign-language translation, intermediate representations of sign-facing systems, weak supervision in data-limited contexts, and syntax-aware reordering in machine translation. Rather than presenting an exhaustive overview of all relevant research, this chapter focuses specifically on prior work that highlights the problem addressed in this thesis and motivates the chosen methodological strategy.

A foundational assumption in this thesis is that sign-language systems do not always need to be formulated as end-to-end predictions. In a low-resource setting, end-to-end learning faces fundamental constraints: target data is rare and expensive to annotate [2], and inconsistent data raises additional obstacles [19]. It therefore pays to break down the task into modular subproblems: each stage of the text-to-sign processing can be considered separately, evaluated, and later connected to the overall pipeline architecture. Accordingly, the present thesis investigates structural reordering of sign-facing sequences as a stand-alone task.

2.2 Sign-Language Translation as a Distinct Structural Problem

Sign-language technology encompasses recognition [17], generation [25], and translation [23]. Despite these differences, the underlying issues of recognition and translation are structurally similar, as sign languages cannot easily be considered visual renderings of spoken-language texts. Bragg et al. [2] claim that sign-language processing requires an interdisciplinary perspective involving linguistics, accessibility research, natural language processing, computer vision, and human-computer interactions. This claim, in turn, suggests that sign-language translation involves structures beyond simple lexical substitution.

Surveys confirm that sign language-to-spoken language and spoken language-to-sign language translation still face data bottlenecks and substantial structural mismatches between spoken and signed representations [7, 9, 23]. For example, De Coster et al. [7] emphasize that representation choices determine what can be learned and what can be evaluated. Farooq et al. [9] systematically review bottlenecks for data and model evaluation. In their survey of traditional rule-based and statistical systems, Núñez-Marcos et al. [23] highlight the impact of dataset limitations on reported results. Finally, Kahlon and Singh [14] provide an historical overview of text-to-sign translation methodologies, reinforcing the point that data-sparse evaluation criteria remain a common issue.

Despite these challenges, neural sequence models hold promise for sign-language problems. Neural sign-language translation was first proposed by Camgoz et al. [3]: a CNN encoder and sequence-to-sequence decoder were combined for the task, and in subsequent work the authors showed that the task could be performed jointly with sign-language recognition using transformer-based architectures [4]. At the same time, however, end-to-end sign-language models depend on datasets and target annotations that are often not readily available in low-resource settings. This thesis explores whether an upstream structure transformation is already valuable, before moving onto target translation and generation.

The use of gloss in sign-language problems is discussed in more detail below.

2.3 Intermediate Representations in Sign-Facing Systems

Intermediate representations are a critical piece of sign-language processing architectures since the conversion from text to visual output is rarely straightforward. In many systems, spoken and written input passes through some symbolic processing step before reaching the final sign-generation stage. For instance, a text-to-gloss-to-sign pipeline was described by Rastgoo et al. [25], and Kahlon and Singh [14] review a range of formats used as intermediaries. This modular approach allows development of linguistic transformation processes separately from animation or visual rendering.

In low-resource settings, an intermediary format facilitates decomposition of sign language problems into manageable subproblems [18, 19]. For example, Moryossef et al. [18] employ gloss augmentation to solve the task, and in subsequent research, the authors explicitly state the unsuitability of full supervision at current data scales [19]. This thesis is inspired by the same principle but applied to a simpler subproblem. Rather than aiming for gloss, SignWriting, or pose as an intermediary format, the goal is to see whether a textual intermediate representation suffices. In contrast to gloss, sign, and pose, this representation is easier to generate automatically and evaluate with text-based measures, as well as directly processible by a conventional sequence-to-sequence architecture.

Relevant literature on the use of various intermediate representations for sign-language problems further reinforces this modularity. For example, Jiang et al. [13] explore translation between spoken and signed languages using SignWriting as an intermediary form, and their results confirm that non-gloss forms can be successfully translated as well. De Coster and Dambre [6] find that freezing pretrained written language representations can aid sign language processing, supporting the general argument that written language modeling can contribute to sign language problems

despite the final sign-language objective. Gómez et al. [10] demonstrate the effectiveness of transformers in translating from text to gloss by explicitly incorporating syntactic information: although their target representation is gloss, like in the previous cases, the study also shows that syntactic considerations are important for low-resource sign-language problems.

2.4 Weak Supervision and Data Scarcity in Low-Resource Settings

Annotated datasets for sign languages tend to be small and expensive to construct. This aspect of sign language research is documented in both interdisciplinary [2] and systematic [7] overviews of the field. Annotation processes often involve gloss alignment, segmentation, translation, notation, and pose labeling, which makes supervised sign language learning a challenge in many scenarios. As a result, low-resource sign language research often relies on weakly supervised methods.

Generally speaking, weak supervision and pseudo-labeling are increasingly popular paradigms in natural language processing. Weak supervision in the form of programmatic label combination was shown to produce sufficient supervision even in the absence of manually created labels [26]. Zhang et al. [28] provide a comprehensive overview of conditions under which pseudo-labels are beneficial for training. Along a related line, Hinton et al. [12] demonstrated that knowledge distillation could make the training process less reliant on strong labels. In this context, Kim and Rush [16] generalized distillation to neural machine translation. This thesis makes use of both ideas: pseudo-labels generated by a rule-based structured teacher enable weak supervision.

As a result, in addition to being weak, supervision in this thesis is also structured rather than manually inspected. High accuracy on the task does not indicate acquisition of sophisticated sign language knowledge, but it does mean that a model learns to replicate a structured teacher that produces sufficiently consistent output to

be useful as intermediate supervision. Additionally, weakly supervised models in the current setup are highly interpretable: each target can be explained as a rule-based computation rather than as black-box network behavior.

2.5 Syntax-Aware Reordering and Structural Transfer

Sign language translation is one motivation behind the present thesis, but there is also a direct connection to the area of syntax-aware reordering in machine translation. Bisazza and Federico [1] offer a survey of word reordering techniques in statistical machine translation, both string-based and tree-based. The importance of reordering is highlighted in their work, with particular attention paid to the role of reordering in accounting for differences between the source and target languages’ grammars. This insight readily transfers from the spoken-to-spoken to spoken-to-signed scenario.

Prior syntax-aware approaches offer even greater precedent. Collins et al. [5] demonstrated the usefulness of reordering to improve German-English statistical machine translation. Neubig et al. [21] proposed training a parser from parallel corpus data to perform optimal reordering. With the advent of neural machine translation, the role of pre-ordering was re-evaluated. Pre-ordering proved effective when working with distant language pairs [29], and even with transformers [15]. Explicit reordering continued to help even extremely low-resource languages [20]. Dependency information improved attention-based models [24].

Taken together, these results confirm that syntax-aware reordering remains a helpful preprocessing step for machine translation, even after the transition to neural architectures. The present thesis proposes a similar approach, but with a different target: rather than producing a natural sentence in a target language, this thesis aims to produce a sequence useful for future sign-facing operations. The novelty of this proposal, then, is not that syntax can be used to reorder input sentences, but that syntax-aware reordering can be used as a sign-oriented intermediate representation

learning task under weak supervision. Thus, this thesis builds a bridge between two distinct fields of study: sign language translation, which requires structural manipulation but suffers from lack of resources, and machine translation, where reordering is well-studied but generally yields spoken-language outputs.

2.6 Syntactic Resources: Universal Dependencies

The linguistic basis of the current thesis is the Universal Dependencies (UD) framework, as described in detail by De Marneffe et al. [8] and Nivre et al. [22]. UD is a joint community project dedicated to developing cross-linguistically consistent treebank annotation for a large set of languages. The framework allows treebank annotation at the token level to include information about word forms, lemmas, part-of-speech tags, morphological features, and labeled dependency relations.

For the purpose of this thesis, the Russian SynTagRus treebank is utilized. This decision is justified in Section 3.4.1. UD is particularly convenient in the context of this research: since the framework is designed cross-linguistically, its methodology is potentially applicable to any language with a UD treebank. In fact, as will be explained later, this is one of the key advantages of UD: it provides opportunities for rule induction without relying on manual inspection.

2.7 Relation to Earlier Pipeline Work

Finally, this thesis is partially motivated by previous system development work in addition to theoretical considerations. Prior sign language pipeline work employed lexical databases, word normalization, pose-format conversion, and animation-based services to generate sign language from spoken language. This approach showed that sign language translation is possible via a multi-step process, but at the same time revealed one critical bottleneck: lexical replacement and pose concatenation are insufficient to account for structural differences between spoken and signed language.

As such, the current thesis can be viewed as continuing along the trajectory of pre-

vious work. Rather than designing an entire new pipeline, it focuses on strengthening one of its constituent elements—a structural reordering component.

Chapter 3

Methodology

3.1 Introduction

In this chapter, we introduce the methodological background of the thesis. We start by formally defining the task and describing the creation of the dataset and pseudo-targets. Next, we describe the neural models that learn to predict the intermediate order of spoken-language sentences. Our proposed pipeline is meant to be modular, i.e., starting from the representation of a spoken-language sentence through its dependency structure, moving on to transforming that sentence into pseudo-target IR through a deterministic rule system, and finally using these targets to train a text-to-text sequence-to-sequence model predicting IR from plain text.

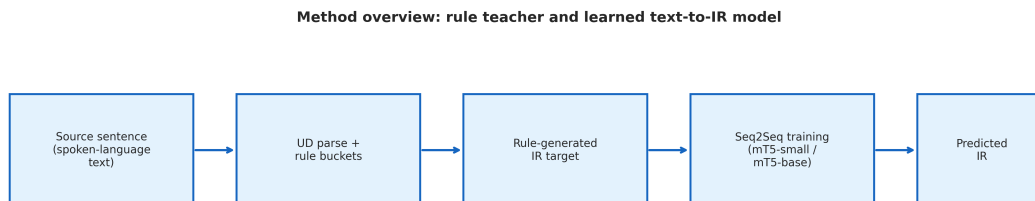


Figure 3-1: The basic idea behind the proposed pipeline. From a spoken-language sentence annotated with dependencies, we create a pseudo-target IR through a rule system, which serves as a label for a text-to-text machine-learning system.

3.2 Notations and Terminology

Let $x = (x_1, x_2, \dots, x_n)$ be a spoken-language sentence in the form of a sequence of tokens. In the present thesis, x usually corresponds to a spoken-language sentence in the UD Russian SynTagRus dependency treebank. Let $y = (y_1, y_2, \dots, y_n)$ be the output reordered sequence. This output sequence contains almost exactly the same lexical material as the input sequence but in a rearranged order. Hence, our task is not lexical replacement but structural reordering.

The central notion of the thesis is the *Intermediate Representation* (IR), which we have introduced in Section 1.3. In this thesis, the IR is a sequence of text items that stands between the raw spoken-language input and further sign-processing operations. This representation is a textual, rather than visual or symbolic in a wider linguistic sense, reorder.

The second crucial concept is *weak supervision*. The targets y are not provided by expert annotators in our case but are created automatically by a deterministic rule system. Consequently, we refer to them as pseudo-targets or pseudo-labels, and our task is to make our model learn to imitate these targets.

3.3 Task Definition

Our task, formalized in this thesis, is as follows: given a spoken-language sentence x , predict a reordered sequence y that resembles the output of a deterministic rule system representing an intermediate order of signs. This is a text-to-text transformation task. The output remains in text format, but its token order becomes changed in order to be suitable for downstream sign-finding tasks.

We define the task as narrowly as possible to achieve two aims: (i) to avoid claims beyond what we intend to study, namely the learnability of structural bottlenecks in sign-finding; (ii) to provide a benchmark for testing whether such bottlenecks can be learned prior to the full task of generating sign-language.

3.4 Dataset Construction and Filtering

3.4.1 Syntactic Backbone

The syntactic resource that we use in this thesis is UD Russian SynTagRus. The corpus provides us with token forms, their lemmas, universal POS tags, dependency relations, head indices, and morphological features. All these syntactic components allow us to define deterministic rules in terms of syntactic roles like subject, object, oblique modifier, adverbial modifier, root predicate, etc.

It is important to emphasize that we do not imply that the induced IR will necessarily be translated into any particular sign language but simply try to isolate the structural bottleneck to be solved first as a learning problem. To this end, we choose SynTagRus because of its syntactically rich annotation allowing rule induction and controlled experiments with texts. In other words, our syntactic corpus is to be used as a structural backbone and not a substitute for sign-language data.

As for the statistics of the used dataset, 69,631 sentences were examined, 6,000 sentences were kept, and the split is made up of 4,800 training examples, 600 validation examples, and 600 test examples.

3.4.2 Quantitative Filtering Pipeline

Since the syntactic corpus contains many types of sentences where internal reordering is hard to achieve, e.g., sentences containing many clauses and subordinated clauses, we filtered out many of the sentences. As our objective is to first prove that we can induce IR mapping and then learn it, the dataset is filtered as strictly as possible. Table 3.1 details the number of sentences removed at each stage of the filtering pipeline.

Filter stage	Sentences removed	Sentences kept
Raw corpus	—	69,631
Length filter (3–20 tokens)	16,926 (24.3%)	52,705
Multi-root removal	0	52,705
Question removal	1	52,704
Subordination removal	30,379 (43.6%)	22,325
Complex clause removal	461	21,864
Random subsample	15,864	6,000
Final dataset	—	6,000

Table 3.1: Detailed filtering statistics.

First of all, one can note that the most common filter is the removal of subordination, filtering out 43.6% of sentences that remain after the length filter. This indicates that complex sentences with embeddings are rather common in the syntactic resource under consideration, proving that the proposed simplification makes sense. Therefore, it is not only a way of reducing the amount of data, but also a method of defining the task itself by selecting certain sentences that can be reordered reasonably.

There may arise some doubts concerning the appropriateness of limiting the benchmark by simple declarative sentences. There are several arguments in favor of such a choice. First of all, declarative sentences prevail in situations related to conveying information (such as news, public statements, etc.), whereas questions and commands are associated with a qualitatively different grammatical structure (involving non-manual markers or peculiar prosody). Second, multi-clause sentences with nested structures may lead to the appearance of ambiguity affecting the quality of generated pseudo-targets. This rule set is applicable to a single predicate declarative clause with clearly defined reorderings, while more complex structures would make this process ambiguous. However, the proposed selection still includes various syntactic structures. The sentence length varies from 3 to 20 words. In addition, the set includes intransitive, transitive, ditransitive, and nominal-predicate sentences.

Thus, we ended up with 6,000 sentences, split into 4,800 sentences for training, 600 sentences for validation, and 600 sentences for testing. The split is stratified by buckets of sentence length. Stratified split ensures similar conditions for each experimental setup since all three sets include sentences of all possible lengths and structures. Our task is to conduct controlled experiments with a limited amount of sentences. Therefore, our selection of sentences is motivated by the task. We want to provide a solid basis for the investigation of learnability, stability, and structural behavior of our method.

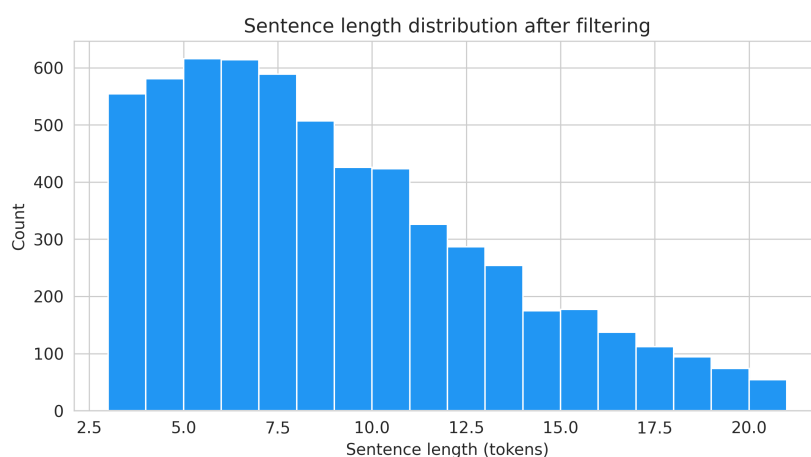


Figure 3-2: Distribution of sentences by length after filtering.

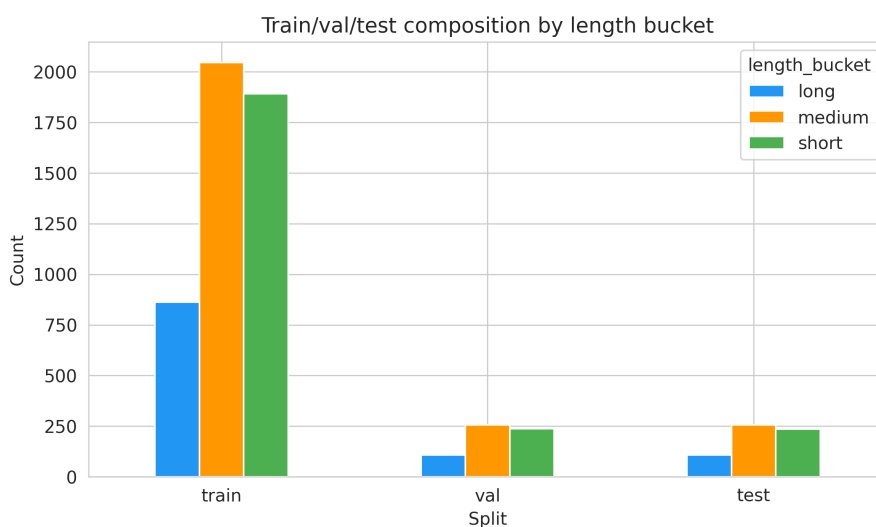


Figure 3-3: Split composition by sentence length buckets.

3.5 Weakly Supervised Target Induction

3.5.1 Motivation

Complete gold sign-oriented reorderings are hard to come by, particularly in low-resource environments. To sidestep this problem, pseudo-labels generated automatically were used instead. The basic idea is that the dependency tree already contains enough structural information to enable a rule-based approximation of a sign-oriented token order. As long as the induced rules produce consistent and stable pseudo-targets, these will serve as the basis for training a neural model.

It is important to view this procedure as structural bootstrapping. The goal of the rule system is to approximate sign language structure, but not to fully model the complex syntactic properties of the underlying grammar. Instead, the aim of the rule system is to generate a proxy target distribution that is structurally explicit and reproducible, yet consistent and stable enough to allow learning.

3.5.2 Rule Design

The rules are not intended to model all aspects of sign language grammar. Rather, they define a target distribution that is interpretable and easy to reproduce from dependency syntax while being stable enough to enable supervised learning. Broadly speaking, the reordering rules favor earlier placement of major arguments (including negation) and predicate delay.

Specifically, the rules try to satisfy the following desiderata:

- Time expressions appear early in the sequence (temporal fronting)
- Subjects are placed before predicates
- Object arguments precede predicates
- Negation is kept close to verbal elements
- Auxiliaries are down-prioritized compared to lexical predicates

- Left-to-right order is preserved within rule buckets

Rule	Intuition	Signals	Action
Temporal fronting	Time expressions appear early in the sequence	advmod, obl + temporal cues	Move to sentence initial region
Subject before predicate	Maintain prominent argument before clause final predicate	nsubj	Place before other arguments and verb block
Object before predicate	Facilitate predicate-final structure	obj, iobj	Place before verb block
Negation near predicate	Maintain proximity to verbal elements	Negation particle + advmod	Place immediately before verb block
Auxiliary de-prioritization	Lexical verbs have higher semantic weight	AUX, cop	Place after the main verb
Modifier handling	Prevent spreading of nominal modifiers	amod, det, nummod, PRON	Place before verb block
Tie breaking	Prevent arbitrary reordering within buckets	original token id	Maintain left-to-right order

Table 3.2: Dependency-based rule system used to generate pseudo-targets.

The above list strikes a practical compromise between complexity and interpretability. The rules are explicit enough to generate stable and consistent pseudo-targets, yet simple enough to remain interpretable and reproducible. Their goal is

not to approximate all aspects of sign language structure, but to induce a limited set of ordering preferences that could be modeled.

3.5.3 Linguistic Rationale and Scope of the Rules

Each individual rule serves the broader goal of reducing dependency on the original order while creating a sequence more amenable to further sign-facing transformation.

Temporal fronting is justified by the common use of temporal expressions as high salience contextual markers. By placing them early in the sequence, one reduces reliance on the original sentence order while creating a framework to introduce the predicate later.

Rules about placing subjects and objects before predicates reflect the overall tendency toward delaying the verbal region and prioritizing arguments over verbal elements. The two rules are particularly important to the induced proxy target distribution since they create an interesting structural transformation and not merely a token shuffling operation.

Placing negation close to the verbal material preserves interpretability, making the pseudo-targets more coherent.

Down-prioritization of auxiliaries is justified by the intuition that lexical verbs carry higher semantic load and need to be placed first to properly express the intended semantics.

Preventing arbitrary scattering of modifiers is essential for creating a consistent target distribution.

Finally, tie breaking rules based on original token order prevent instability caused by arbitrary reordering within each bucket.

At the same time, the above rules are deliberately focused only on simple token re-ordering and ignore such linguistic phenomena as discourse organization, non-manual markers, classifiers, and more complex properties of sign language grammar. The point here is simply to induce an interpretable structural transformation that can be reliably generated from dependency syntax and learned as a separate task.

3.5.4 Examples

This section provides three worked examples tracing the rule system from source sentence to pseudo-target. Each example shows the token attributes and the resulting reordering.

Example A — Auxiliary deprioritization

Token	id	UPOS	deprel	Rule bucket
Всё	1	PRON	nsubj	subj
было	2	AUX	aux	auxs
тщетно	3	ADV	root	verbs

Source: Всё было тщетно → Pseudo-target: Всё тщетно было. The auxiliary было is deprioritized to the `auxs` bucket, placing it after the lexical predicate.

Example B — Temporal fronting and SOV reordering

Token	id	UPOS	deprel	Rule bucket
Вчера	1	ADV	advmod	tadv
студент	2	NOUN	nsubj	subj
сдал	3	VERB	root	verbs
экзамен	4	NOUN	obj	objs

Source: Вчера студент сдал экзамен → Pseudo-target: Вчера студент экзамен сдал. The temporal adverb Вчера is promoted to the `tadv` bucket (sentence-initial); the object precedes the verb, producing SOV order.

Example C — Negation placement

Token	id	UPOS	deprel	Rule bucket
Он	1	PRON	nsubj	subj
не	2	PART	advmod	negs
читал	3	VERB	root	verbs
книгу	4	NOUN	obj	objs

Source: Он не читал книгу $\rightarrow$ Pseudo-target: Он книгу не читал. The negation particle не is placed in the **negs** bucket immediately before the verb, while the object книгу precedes both.

These three examples cover the main rule types — auxiliary deprioritization, temporal fronting, and negation-predicate adjacency — and confirm that the rule system applies consistent, interpretable structural preferences across all 6,000 sentences in the dataset.

3.6 Neural Reordering Models

3.6.1 Why a Neural Model Is Needed

The first and foremost question concerns why a neural model is needed at all given that the rules were already described. There are three good reasons for that.

First, a neural model makes the application more robust to input variation. While a rule system needs dependency parsing to run inference, a learned sequence generator can map text input to IR in a single step.

Second, the neural sequence model will be more robust against dependency parsing mistakes and other variations not considered by the rules. As we saw previously, some examples required manual correction of the parser output to get the correct pseudo-labels. This is something that a neural model can mitigate by smoothing over such noise.

Third, having learned an IR mapping, it is possible to refine the learned model by incorporating smaller datasets of validated samples or by combining manual labeling

and target approximation.

In effect, the neural model represents an attempt to learn the current rule system and can be improved further by refining the input data or adding more sophisticated target definition.

3.6.2 Model Family and Capacity Comparison

For this experiment, the chosen family of models is mT5 [27], a multilingual extension of T5 (Text-to-Text Transfer Transformer) model. mT5 is trained using a variant of the Common Crawl corpus known as mC4, with the latter covering 101 languages, including Russian. The architecture follows a common encoder-decoder design: the sequence input is encoded using self-attention and feed-forward layers, yielding a representation of the input sequence; this representation is subsequently decoded via cross-attention and autoregressive generation of the output sequence.

The reason to choose mT5 is threefold. First, its text-to-text paradigm perfectly fits the task. Since IR is essentially defined as a token reordering task, mT5’s generic text-to-text paradigm offers a straightforward way to solve the task without architectural modification. Second, mT5 covers multilingual text and includes support for Russian language, which matters since source sentences are taken from the SynTa-gRus treebank. Finally, it allows to perform a controlled model-capacity comparison, which would not be possible with differently trained models.

To evaluate different levels of model capacity, two sizes of mT5 are used: small (~ 300 million parameters) and base (~ 580 million parameters). While both models are built from the same architecture with identical hyperparameters, pretraining, and tokenization, they differ only in terms of architecture specifics (number of layers, number of attention heads).

In the previous iterations of the experiment, only mT5-small was used. The reason to include mT5-base in the experiment is to rule out the critique that the thesis relies on a single model and therefore conflates task difficulty with model limitations. By including both mT5-small and mT5-base, one can demonstrate if adding capacity improves performance. The experiments below aim to test this.

3.6.3 Training Configuration

Both models used in the experiments were trained using aligned hyperparameters to facilitate fair comparison: five epochs, maximum lengths of source/target 96 tokens, learning rate 3×10^{-4} , weight decay 0.01, beam search decoding with beam width of 1. The experiments using the revised model also fixed several important technical issues: aligned ablation experiment hyperparameters, correct propagation of the seed value in Hugging Face configuration file, explicit GPU memory cleanup between runs, disabled mixed precision training for better stability, force single-GPU execution, and reduced batch size with gradient accumulation for larger model.

The above considerations regarding alignment of configurations is extremely important for obtaining reliable results. Indeed, any discrepancies in training configuration and hyperparameters may have an unintended influence on training outcomes, making it impossible to isolate factors contributing to the observed performance differences.

3.6.4 Why These Metrics Are Appropriate

It is worth noting that the combination of the five metrics chosen constitutes a complementary evaluation suite which, taken together, evaluates performance on three distinct criteria mentioned in the thesis’s claims: surface similarity, preservation of contents, and structural accuracy. No single metric addresses all three and, therefore, the specific choice of the metrics used is carefully selected to avoid being too reliant on any one metric.

Firstly, BLEU is included for comparative purposes in order to allow evaluation against previous works related to sequence transduction and machine translation reordering. While BLEU was designed primarily for translation tasks, the use of this metric for the present problem is justified since the objective of this research is the structural reordering of the input string. The n-gram precision nature of the metric means that it does not tolerate any kind of transposition in the output and hence is a strict metric for structural similarity. However, BLEU alone cannot distinguish

between errors in structure and in content, which is why other metrics are needed.

ChrF (character-level F-score) was used to evaluate the ability of the model to account for morphological variability of Russian. In Russian, lemmas are inflected according to various criteria such as case, number, gender, and aspect. This means that there can be significant morphological variance in the surface forms of one and the same lemma. While BLEU will treat all those different forms as distinct unigrams, thus severely punishing the model, chrF, as character-based, will be less affected by morphological differences.

Exact match serves as a strict binary sentence-level criterion for assessing whether or not the output produced is structurally similar to the pseudo-target. While EM rates are fairly low (0.27–0.35), this nonetheless demonstrates that the task is non-trivial and the increase of mT5-small to mT5-base (from 0.27 to 0.35) indicates a true improvement in structural completeness.

The fifth metric, token accuracy, evaluates position-wise overlap between the actual and expected output and, unlike BLEU, is not dependent on n-gram contiguity. This means that TokAcc can detect whether or not the model gets the structure partially right.

Finally, Kendall’s τ , perhaps the most relevant metric in the context of reordering tasks, evaluates the correlation between the token orders in the predicted and target outputs. The correlation rate of 0.86 seen with mT5-base proves that, despite some minor content errors which lower BLEU scores, the model still captures the overall structure of the pseudo-target. Thus, Kendall’s τ allows one to verify the claim of this thesis that the model learns to reorder tokens in a manner preferred by the rule system.

Together, the combination of the five metrics partitions the space of evaluation such that a good performance on all of them implies that the model can generate the output with tokens that are structurally, semantically, and surface-wise correct (EM, BLEU, chrF), in the right order (τ). The lack of human-based evaluation is intentional due to the nature of the pseudo-target and is covered in Section 5.5.

3.6.5 Reproducibility Considerations

When designing experiments, I aimed to make results reproducible whenever possible. Particularly, for the main experiment, the train/dev/test split was created using a fixed seed and reused consistently across experiments. In case of mT5-small ablations, several seed values were used to estimate standard deviation from the results. For the new ablation, hyperparameters for the main experiment and the ablation were aligned to ensure consistent training configuration. Preprocessing and model training were fully deterministic, with outputs stored to disk. Intermediate predictions and logging information were exported and could be inspected post-experiment.

Chapter 4

Experiments and Results

4.1 Experiments Overview

The experiments seek to answer four questions. First, is a neural model able to learn the rule-derived sign-oriented target distribution? Second, does the learned model surpass the baseline where the sentence is left in its original spoken-language order? Third, does model capacity influence the task performance? Fourth, which parts of the target generation process have the highest importance?

In order to answer these questions, the thesis reports a source-order baseline, a rule baseline, results for `mT5-small` and `mT5-base`, multiseed variance evaluation, performance by sentence-length bucket, epoch-wise learning dynamics, quantitative error analysis, qualitative samples, and a comprehensive ablation study. Together, these experiments are designed to assess both general imitation performance as well as structure reproduction, result robustness, and the leading failure modes of the trained models.

4.2 Baselines and Evaluation Protocol

The source-order baseline maintains the sentence in its original form. The baseline serves as a measure of the extent to which the source language form deviates from the sign-induced target form. The rule baseline is the deterministic system of rules

used for target generation. As the pseudo-targets are produced by this system, the rule baseline serves as a reference ceiling on rule imitation.

Evaluation involves overlapping, character-level, and order-aware metrics. BLEU serves as a widely accepted benchmark metric for sequence generation, while chrF is added as a complementary character-level overlap metric that works well with morphologically rich languages such as Russian. Exact match and token accuracy offer additional views of the accuracy of generation, stricter for sentences as whole and for particular positions within a sentence respectively. Finally, as the task seeks to learn to imitate a certain ordering pattern, Kendall’s τ offers a valuable measure of structure replication. In addition, structural rates such as verb-final alignment and fronting of arguments prior to verbs are used as a way to confirm whether learned ordering patterns actually hold rather than merely imitating surfaces.

4.3 Main Test-Set Results

Model	BLEU	chrF	EM	TokAcc	τ
SourceOrder	14.16	73.36	0.04	0.23	0.20
mT5-small (3 seeds)	52.57 ± 0.49	80.86 ± 0.44	0.27	0.52	0.80
mT5-base	63.90	87.53	0.35	0.62	0.86
RuleBaseline	100.00	100.00	1.00	1.00	1.00

Table 4.1: Main test-set results. **mT5-small** reports mean $\pm$ standard deviation over three seeds.

As one might expect, the source-order baseline shows a poor performance, confirming that there is significant distance between the source sequence and the induced target sequence. Both learned models achieve significantly higher performance compared to the source-order baseline on all reported metrics. The gap between **mT5-small** and **mT5-base** is rather large, over 11 BLEU points, and will be addressed in detail in the following section.

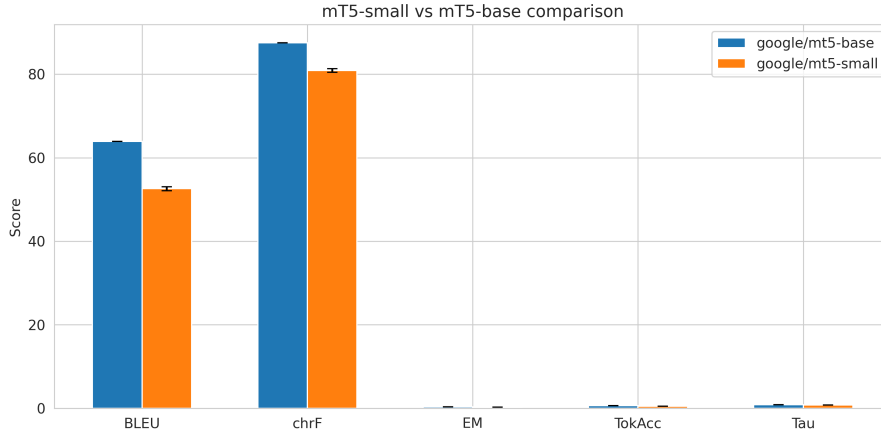


Figure 4-1: mT5-small vs. mT5-base on the main test-set evaluation metrics.

4.4 Effect of Model Capacity

The comparison between `mT5-small` and `mT5-base` constitutes perhaps the most critical finding of the current work. Generally, when working on tasks involving text transformations, increasing model size beyond a certain point leads to diminishing returns due to quick saturation. This is not the case for the presented task. The improvement in BLEU scores over the same architecture with doubled capacity suggests that the induced sign-oriented target distribution still has plenty of complexity and structure to be learned.

In addition, `mT5-base` outperforms the small variant across all evaluation metrics, not only in BLEU. This implies that `mT5-base` has a better understanding of the induced target order, retains more source information, and reduces generation errors. This is an important finding, as it shifts our understanding of the problem domain, implying the existence of sufficient regularities and complexity in the target language that may be effectively learned by a stronger model.

4.5 Sensitivity to Random Seed Initialization

Multiseed training of `mT5-small` with three different random initializations allows us to gauge the degree of variability associated with randomly initialized models. The variance of the reported BLEU scores across seeds is 0.49, while the variance in chrF is 0.44.

The low standard deviations of `mT5-small` performance on both BLEU and chrF indicate that the learning process is relatively insensitive to random seed initialization. This finding is especially important in the context of weak supervision of the task, as it implies high stability and repeatability of results despite minimal information about the target space.

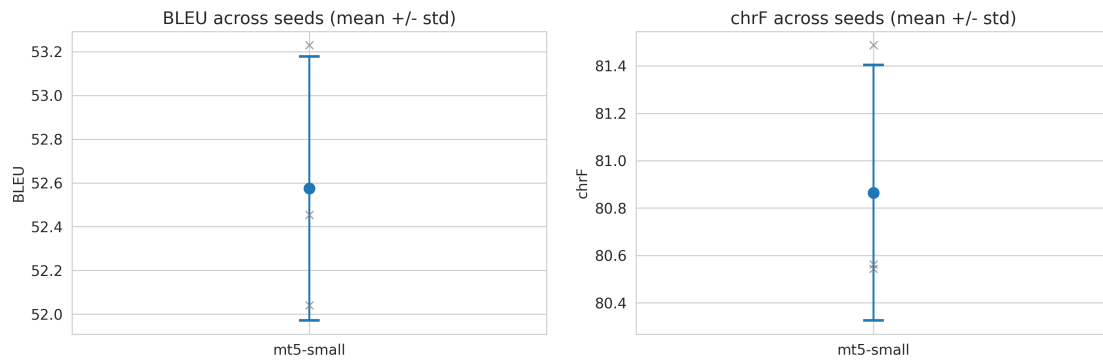


Figure 4-2: Visualizations of BLEU and chrF variability across `mT5-small` seeds.

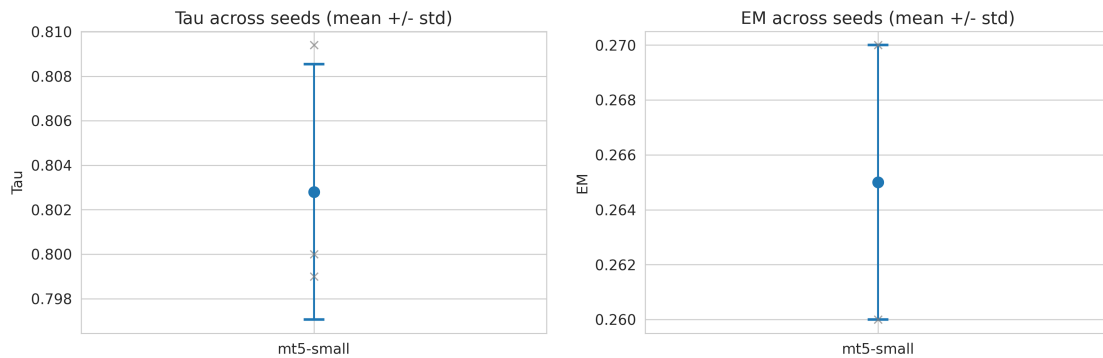


Figure 4-3: Visualizations of Kendall's τ and exact match variability across `mT5-small` seeds.

4.6 Sentence Length Analysis

Bucket	BLEU	chrF	EM	TokAcc	τ
Short	70.02	89.07	0.53	0.72	0.80
Medium	53.70	82.32	0.14	0.46	0.81
Long	45.26	75.96	0.01	0.25	0.78

Table 4.2: Performance by sentence-length bucket for `mT5-small` (best seed).

Performance deteriorates as the length of the sentence grows, which is natural, since longer sentences tend to contain more modifiers, oblique elements, coordination, and internal nominal structures. Short sentences are much easier to reorder precisely, while long sentences are rarely matched precisely. Nevertheless, even in long sentences, the value of τ remains fairly high, suggesting that the model successfully learns the general structural order of elements within a sentence despite occasional mistakes in exact ordering.

It is important to note that, although the decrease in performance is quite noticeable, it does not imply a complete failure on the part of the model. Instead, it demonstrates that the model learns to maintain a general structural order while failing to preserve the full content and generate all tokens in their proper order.

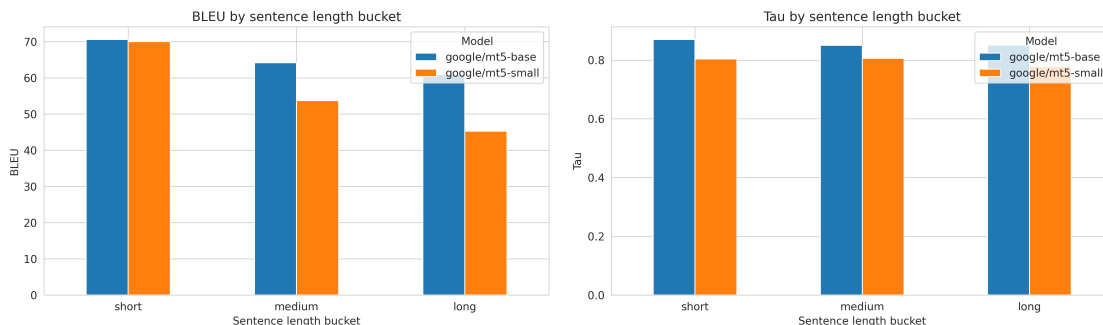


Figure 4-4: Performance by sentence length buckets for BLEU and τ .

4.7 Learning Dynamics

The epoch-wise graphs provide an insight into model learning. For `mT5-small`, validation BLEU starts from around 21 and increases to 53 by epoch 5. At the same time, Kendall's τ increases from roughly 0.61 to 0.80. The model with doubled capacity learns faster and achieves a higher maximum performance.

The learning curves suggest stable and healthy training for both variants of the model. The larger variant is trained to reach higher levels of validation BLEU score and, unlike `mT5-small`, continues to converge. While both models demonstrate a growth in τ , overfitting does not seem to occur in either case, at least not to a detrimental degree.

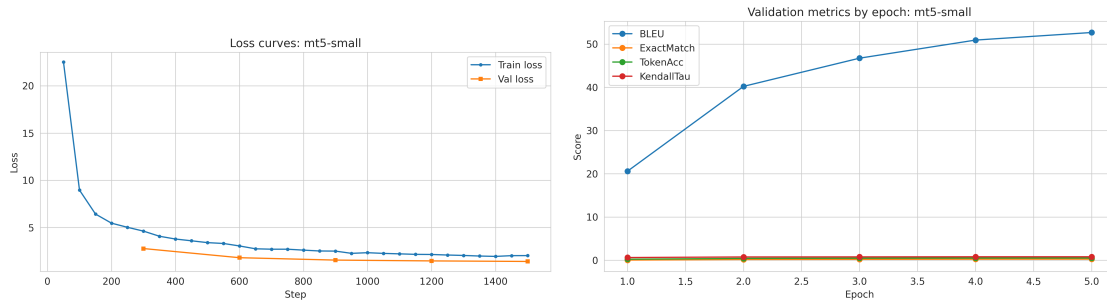


Figure 4-5: Training loss and epoch-wise generation metrics for `mT5-small`.

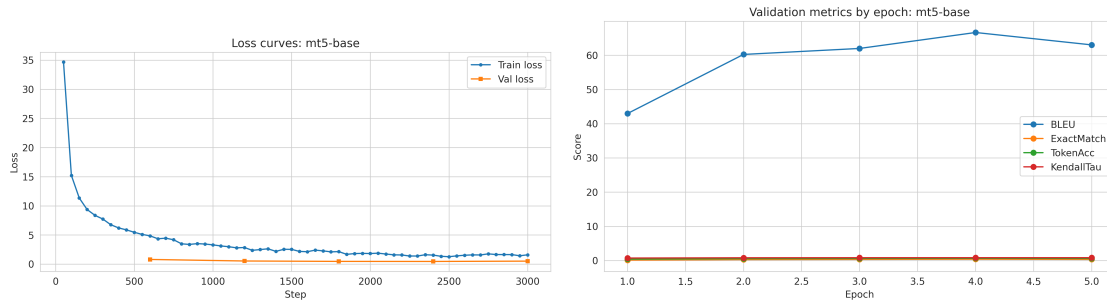


Figure 4-6: Training loss and epoch-wise generation metrics for `mT5-base`.

4.8 Error Analysis

Four types of errors were measured using the error analysis framework: repetition (repeated appearance of a source token within the predicted sequence), omission (absence of a source token from the predicted sequence), hallucination (an extra token appearing in the predicted sequence), and length mismatch (predicted sequence length being different from the target length).

Bucket	Repetition	Omission	Length mismatch	Hallucination
Overall	0.290	0.462	0.442	0.053
Short	0.123	0.144	0.178	0.025
Medium	0.309	0.578	0.523	0.055
Long	0.611	0.880	0.824	0.111

Table 4.3: Error rates for `mT5-small` on the test set, overall and by sentence-length bucket.

As one might expect, length is the most critical factor here. Omission rates increase from 14.4% on short sentences to 88.0% on long sentences. Similarly, repetition rises from 12.3% to 61.1%. The rates of hallucination remain rather low, suggesting that the model rarely generates tokens outside the source vocabulary. Thus, the main difficulty seems to lie in preserving the entire list of source tokens along with the correct order.

This error analysis proves to be important, as it demonstrates that the model struggles with content preservation rather than uncontrolled generation.

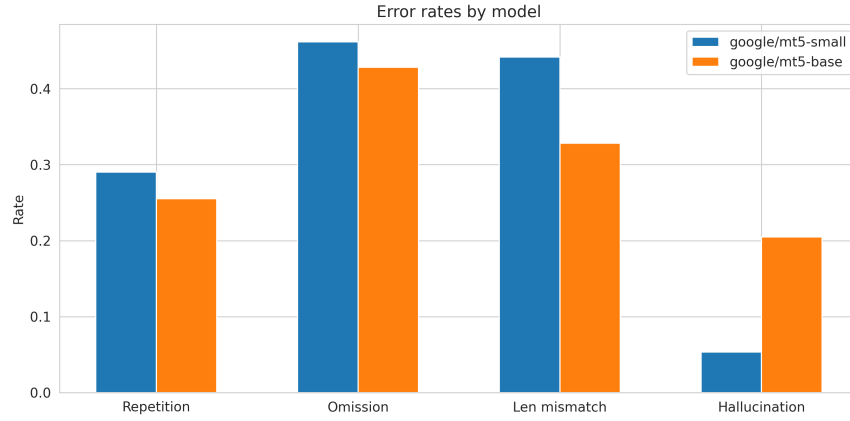


Figure 4-7: Error rates for repetition, omission, hallucination, and length mismatch.

4.9 Sample Examples

To demonstrate the behavior of the learned model in detail, the thesis considers four test-set examples, contrasting them with source sequences and rule-produced targets.

Example 1 (short, 3 tokens — both models match exactly)

Source Все было тщетно

Target Все тщетно было

mT5-small Все тщетно было

mT5-base Все тщетно было

Example 2 (medium, 5 tokens — mT5-base matches exactly)

Source Какие уж тут советские свадьбы

Target Какие советские уж тут свадьбы

mT5-small свадьбы Какие советские уж тут

mT5-base Какие советские уж тут свадьбы

Example 3 (medium, 12 tokens — mT5-small omits one token)

Source Это связано с проведением шествия митинга и концерта в поддержку
военной реформы

Target Это проведением военной связано и с шествия митинга концерта в
поддержку реформы

mT5-small Это проведением военной связано и с митинга концерта в поддержку
реформы

mT5-base Это проведением военной связано и с шествия митинга концерта в
поддержку реформы

Example 4 (long, 10 tokens — both models omit one token)

Source Во временные промежутки между снимками возможно накопление
ледником критической массы

Target накопление промежутки временные критической возможно Во между
снимками ледником массы

mT5-small накопление промежутки временные критической возможно Во между
снимками массы

In example 1, the model successfully reproduces target ordering, delaying the predicate (было) until the end of the sentence. In example 2, **mT5-base** exactly reproduces the target ordering, while **mT5-small** fails and erroneously fronts the content-bearing noun свадьбы. In example 3, **mT5-small** omits a token шествия while **mT5-base** manages to retain the full set of tokens and reproduces target ordering. In example 4, both models successfully capture the broad ordering pattern but miss one content-bearing token ледником.

These examples are consistent with the observations made during error analysis: omission emerges as the leading type of error, which worsens with sentence length.

4.10 Ablation Study

Finally, an ablation study seeks to find out whether the presented results depend on any of the data preprocessing and target generation choices or can be attributed to the overall coherence of the target language distribution. For ablations, the same hyperparameters are used as in the main experiment, including learning rate and the size of the training sample, in order to attribute observed differences to the effect of the ablated variable.

Variant	BLEU	chrF	τ	EM
clean/surface/fullrules	53.23	81.49	0.80	0.27
clean/surface/keep-punct	58.82	82.18	0.87	0.27
clean/surface/no-aux-deprior	54.34	81.88	0.79	0.27
clean/surface/no-advmod-front	54.16	81.92	0.81	0.27
clean/lemma/fullrules	37.47	73.91	0.61	0.06
broad/surface/fullrules	48.28	79.18	0.78	0.01

Table 4.5: Ablation study results, aligned with the main experiment.

Several key findings emerge. First, the presence of punctuation marks contributes to performance, improving both BLEU and rank preservation. Second, dataset clean-

liness matters greatly: the broader and noisier version of the dataset achieves noticeably lower BLEU score and rank preservation. Third, lemmas produce substantially worse results compared to surface forms, indicating that the introduction of lemmas makes the problem harder to learn.

By contrast, removing particular rule components such as auxiliary deprioritization or adverbial fronting shows moderate effects. Overall, it seems that coherence of the induced target distribution plays a much greater role in the performance than any particular rule used to induce it.

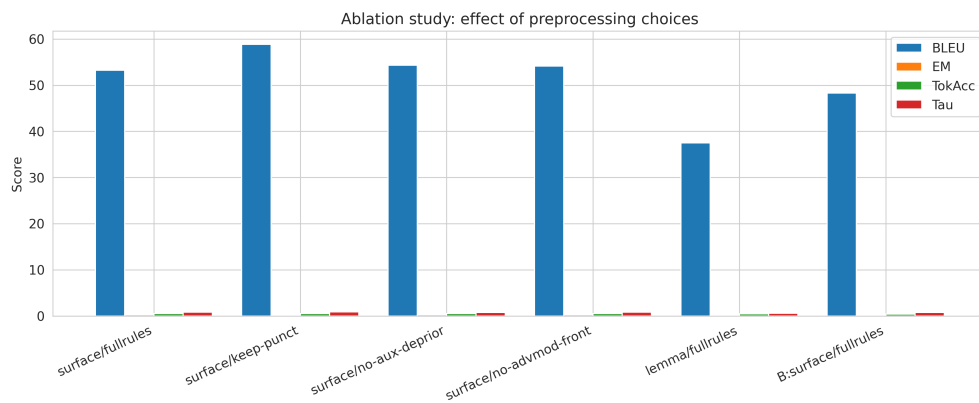


Figure 4-8: Comparative analysis of different preprocessing strategies and target representations.

Chapter 5

Discussion and Conclusion

5.1 Summary of Findings

The key finding of the thesis is that sign-oriented intermediate reordering, a structurally-defined subtask in the broader text-to-sign processing pipeline, can be isolated, learned using weak supervision, and tested independently. Using dependency relations extracted from UD Russian SynTagRus to define weak supervision, the work demonstrates that a sequence-to-sequence multilingual transformer can learn a significant portion of the induced mapping and perform significantly better than keeping the source-order sentence intact.

Several key observations follow from the results presented here. First, the reordering task is learnable. The large difference in performance between the source-order baseline and the learned models suggests that the pseudo-targets do not represent some random space, but rather a stable set of structural characteristics that can be approximated by neural models.

Second, model capacity matters. The 11-point BLEU score gain from `mT5-small` to `mT5-base` cannot be considered a marginal difference. It shows that the learning capacity of the task has not been reached on the level of `mT5-small` and strongly suggests that it still has room for progress. This is one of the most important findings of the revised thesis as it alters the interpretation of the task itself: the current induced target still contains regularities which a stronger model can take advantage of.

Third, the learned target mapping is robust to model initialization. Multi-seed analysis for `mT5-small` shows little variation and suggests that the reported results are not due to an especially fortunate run of the training process. Such result stability is crucial since it enhances the validity of the methodology used by the thesis.

Fourth, failure modes are systematic rather than stochastic. The detailed error analysis shows that repetition and omission occur much more often than hallucination and tend to grow sharply as the sentence length increases. In particular, omission becomes particularly widespread in longer sentences.

Finally, the ablation study shows that dataset quality matters more than individual rule choice. The clean subset proves much easier to learn than the original one; additionally, punctuation retention appears to be helpful for learning. This suggests that the success of the thesis relies less on an optimal rule specification and more on the overall coherence of the induced pseudo-target space.

5.2 Interpretation of the Learned Models

Given that the targets in this thesis are deterministic pseudo-labels derived from a rule system, the learned sequence-to-sequence models should be understood as an approximation of the rule system itself rather than recovered sign language grammatical structure. While high BLEU, chrF, or Kendall’s τ scores suggest a good match between the model predictions and the pseudo-target, they do not indicate actual recovery of a genuine sign-language mapping. On the contrary, it means that there exists a stable and structurally well-formed mapping that can be approximated with a sequence-to-sequence transformer.

In addition to providing evidence that the target space is indeed consistent, the result also shows that the problem remains non-trivial enough for model capacity differences to play a role. The strong BLEU score gain from `mT5-small` to `mT5-base` suggests that there are still regularities left in the induced pseudo-target distribution that can be exploited by larger models. This finding is especially important since it suggests that further gains may be achieved not just through more elaborate rules or

cleaner data preprocessing, but through more capable architectures.

However, it is worth pointing out that the current error profile shows the limitations of the learned models. First, despite the strong performance on chrF metric, hallucination remains rare, which is encouraging: the model does not produce tokens not present in the source. Still, omission and repetition become prevalent as the sentence length increases. This suggests that the difficulty of the task lies in preserving content tokens after reordering and not inventing entirely new tokens.

5.3 What the Results Mean for the Proposed IR

While the results of the thesis certainly show that the induced intermediate representation is feasible to learn, they should be interpreted not as direct evidence of good sign-language translation quality but as evidence that it is scientifically valuable.

The current IR is intentionally limited in terms of expressive power. Unlike the proposed sign-language representations in Section 1.2, it does not encode any non-manual information or spatial references, does not involve classifier constructions, and disregards sign-specific phenomena altogether. It is nevertheless a reasonable target in low-resource settings: the learned IR is structurally more informative than the source order and can be induced automatically using UD data. It also allows for independent evaluation with standard BLEU/chrF metrics prior to integration into broader sign-language pipelines.

In terms of linguistic relevance, the current IR is a compromise between pure structural reordering and sign-language grammatical encoding. Since it is induced using dependency relations extracted from the source, it can be seen as an approximation of the underlying sign-language reordering. This approach can be especially relevant for downstream sign-language tasks, where reordering tends to play a key part.

5.4 Relation to the Earlier Pipeline

As noted above, the earlier thesis on Qazaq sign-language translation involves a multi-stage system, including base-form conversion, lexical retrieval, and pose format conversion. The results of that work demonstrate both feasibility and limitation: while the pipeline proved successful for sign facing applications, its reordering module relied on source-language word order, which did not always correspond to sign language structure.

In this context, the current thesis can be seen as a continuation of that previous work. Specifically, it suggests developing a structurally-driven front-end module that would then be integrated into the previously proposed pipeline. Such relationship is quite important since it puts the current theoretical problem into a concrete system context: unlike being a mere reordering task, this intermediate problem has direct practical meaning since it aims to improve the input to downstream stages.

5.5 Limitations

There are several key limitations of the current thesis that should be highlighted to keep the contribution appropriately scoped.

First, the induced target is still an IR rather than a proper sign language target. Although the pseudo-target generation procedure is rule-based, the rules are written not by sign language experts, native signers, or based on gold sign language datasets. This means that learning this distribution proves structural learnability of the target IR, but does not prove that the learned IR is linguistically valid from a sign language perspective.

Second, the thesis uses BLEU/chrF measures to evaluate the task, thus focusing strictly on textual output. This means that it cannot provide direct evidence of any downstream avatar quality, signer comprehension, or user-facing performance.

Third, although UD Russian SynTagRus is known to include diverse sentence structures, the current dataset is intentionally limited to declarative sentences. This

was done to simplify the evaluation procedure but is nonetheless a limitation preventing generalization to complex sentences.

Fourth, since Russian SynTagRus is the most readily available dependency dataset in this context, the experiments rely on Russian syntax, which cannot be directly transferred to Kazakh sign language and Kazakh-Russian sign language. Therefore, additional linguistic challenges may appear in applying this method to those languages.

Fifth, the error analysis shows that omission and repetition rates have a sharp rise with sentence length, increasing from 14.4% and 12.3% in short sentences to 88.0% and 61.1% in long sentences. This is not merely a mistake but a result of the auto-regressive generation process. With increasing length of the output sequence, the model’s attention gets diluted over an increasingly large history of previous generated tokens, leading to decreased attention on source tokens still to be generated. Particularly vulnerable are tokens from non-distinctive rule buckets, such as oblique modifiers, other adverbials, and the general-purpose `others` bucket; these rule buckets group tokens based on their index position, making this relationship more difficult for the model to learn than others. Given an omission, the model tends to repeat a previously generated token rather than recover the missed token. The fact that hallucination rates are relatively small (5.3% for all sentences, and 11.1% for long sentences) supports the idea that this is a failure of coverage, and not vocabulary modeling.

Finally, since the pseudo-target generation procedure is rule-based, the method used cannot be compared to alternative approaches that might generate targets using another kind of weak supervision.

5.6 Future Work

First, the next step in development is performing end-to-end validation and comparing a system without IR reordering, with manually specified rules, and with learned IR generation. This will allow checking whether the new component contributes to the

quality of the sign-facing system.

Second, future work might involve using a manually validated IR set rather than pseudo-targets generated automatically. This additional weak-supervision layer may be used either to refine the rules or fine-tune the models.

Third, and the one most clearly prompted by the issue of coverage failures due to length, future research directions might include developing an architecture which explicitly covers all of the source. This could be achieved by replacing beam search decoding, which is unsupervised, with constrained decoding which forces the output to be a permutation of the source tokens using coverage which ensures that each source token is marked as either having been decoded or not. In addition to this, adding a pointer network/copy mechanism which could only choose from source tokens would eliminate any issues of hallucinations occurring while at the same time giving the model an indication of what source tokens had yet to be generated. Another possibility includes adding a coverage vector in the attention module itself, which is a sum of the attention weights over source tokens, and penalizing training if the coverage of each source token was insufficiently attended. Finally, the use of curriculum learning to train the model first on sentences where the model performs well (BLEU score of 70.02 on short bucket) then gradually training on longer sentences.

Fourth, since the current performance difference between `mT5-small` and `mT5-base` is substantial, experimenting with even larger models, including language domain-specific models, may be promising. It might also be useful to develop more specialized architectures explicitly targeting constrained reordering tasks.

Fifth, as noted above, currently only relatively simple sentences are used. This makes sense from the point of view of evaluating learnability, but future work should try to generalize to more complex cases, including embedded clauses and discourse markers.

Sixth, as explained above, Russian UD dependency parsing is convenient but not directly relevant to the problem domain. Future work may try to apply the same method to UD datasets for Qazaq sign language and Kazakh-Russian sign language.

Finally, exploring alternative methods of generating IR targets is also a valid

future direction, as the current target generation method was chosen for the purpose of easy evaluation rather than optimal task definition.

5.7 Conclusion

By reformulating the practical problem of text-to-sign processing in low-resource sign-language environments, this thesis presents and validates a new structural task: sign-oriented intermediate reordering. Using UD Russian SynTagRus dataset to generate pseudo-targets, the thesis shows that the sequence-to-sequence transformers are capable of learning the induced target distribution and producing structurally similar sequences. Moreover, the learned model proves significantly superior to the source order baseline and shows strong structural correspondence with the pseudo-targets. Multi-seed analysis confirms the stability of the results; quantitative error analysis demonstrates their failure modes.

The main contribution of the thesis is the demonstration that the task of structural reordering of text can be learned using a weak supervision signal, evaluated independently, and eventually incorporated into an end-to-end sign facing pipeline. For low-resource settings where supervised sign language corpora are scarce, such approach can be both feasible and scientifically justified.

Beyond the specific contribution, this thesis advances a research philosophy. Given that full end-to-end supervision can rarely be obtained for sign language problems, progress should not stall. Instead, it can be made via carefully formulated intermediate tasks, structural assumptions, and modular evaluation of the components.

Bibliography

- [1] Arianna Bisazza and Marcello Federico. A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational Linguistics*, 42(2):163–205, 2016.
- [2] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019.
- [3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.
- [4] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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, 2020.
- [5] Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 531–540, 2005.

- [6] Mathieu De Coster and Joni Dambre. Leveraging frozen pretrained written language models for neural sign language translation. *Information*, 13(5):220, 2022.
- [7] Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23(3):1305–1331, 2024.
- [8] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 2021.
- [9] Uzma Farooq, Mohd Shafry Mohd Rahim, Nabeel Sabir, Amir Hussain, and Adnan Abid. Advances in machine translation for sign language: Approaches, limitations, and challenges. *Neural Computing and Applications*, 33(21):14357–14399, 2021.
- [10] Santiago Egea Gómez, Euan McGill, and Horacio Saggion. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 18–27, 2021.
- [11] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. Word reordering rules for pivot language translation between English and Japanese. In *Proceedings of Machine Translation Summit XIV*, 2013.
- [12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [13] Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. Machine translation between spoken languages and signed languages represented in SignWrit-

- ing. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, 2023.
- [14] Navroz Kaur Kahlon and Williamjeet Singh. Machine translation from text to sign language: A systematic review. *Universal Access in the Information Society*, 22(1):1–35, 2023.
 - [15] Yuki Kawara, Chenhui Chu, and Yuki Arase. Preordering encoding on transformer for translation. *arXiv preprint arXiv:2005.13859*, 2020.
 - [16] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, 2016.
 - [17] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.
 - [18] Amit Moryossef, Mathias Müller, Yoav Goldberg, and Sarah Ebling. Data augmentation for sign language gloss translation. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 365–375, 2021.
 - [19] Amit Moryossef, Zifan Jiang, and Sarah Ebling. Generating data for low-resource sign language translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 2023.
 - [20] Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 3868–3873, 2019.
 - [21] Graham Neubig, Taro Watanabe, and Shinsuke Mori. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the*

2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 843–853, 2012.

- [22] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036, 2020.
- [23] Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993, 2023.
- [24] Ru Peng, Tianyong Hao, and Yi Fang. Syntax-aware neural machine translation directed by syntactic dependency degree. *Neural Computing and Applications*, 33:16609–16625, 2021.
- [25] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Sign language production: A review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3451–3461, 2021.
- [26] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.
- [27] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT 2021*, pages 483–498, 2021.
- [28] Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022.

- [29] Yang Zhao, Jiajun Zhang, and Chengqing Zong. Exploiting pre-ordering for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.