

# DGS Coreference Resolution via Prompt Design: An Evaluation

Yuzheng Bao

March 2025

# Contents

<b>1</b>	<b>BACKGROUND</b>	<b>3</b>
1.1	Background and Motivation . . . . .	1
1.1.1	Reasons of Study Sign Language Pronoun Resolution in Large Lan-	
guage Models . . . . .		1
1.1.2	Distinct Nature of Sign Language . . . . .	1
1.1.3	Bridging the Gap with Large Language Models . . . . .	1
<b>2</b>	<b>Large language models</b>	<b>2</b>
2.1	Definition of LLMs . . . . .	2
2.2	Potentials of LLMs in solving sign language coreference problem . . . . .	2
2.3	The advantage: compare with machine learning . . . . .	3
2.4	Prompt Engineering . . . . .	3
2.5	Zero-Shot . . . . .	4
2.5.1	2.6 Few-Shot Prompting . . . . .	5
<b>3</b>	<b>DGS CORPUS</b>	<b>7</b>
3.1	What is DGS? . . . . .	7
3.1.1	Start and End Nodes, and HAND Annotation . . . . .	8
3.1.2	LEXEME_SIGN, Gebärde, and SIGN Annotations . . . . .	8
3.2	Zero-shot coreference resolution in DGS . . . . .	9
3.3	Few-shot coreference resolution in DGS . . . . .	9
<b>4</b>	<b>Experiment</b>	<b>10</b>
4.1	Research Background . . . . .	10
4.1.1	Signed Coreference Resolution Challenge . . . . .	10
4.1.2	Gold Standard Dataset . . . . .	10
4.1.3	English Translations of DGS Data . . . . .	10
4.1.4	Semantic Analysis Framework . . . . .	11
4.1.5	Glossing Conventions . . . . .	11
4.1.6	Few-Shot and Zero-Shot Paradigms . . . . .	11
4.1.7	Empty-Category Syntactic Modeling . . . . .	11

4.1.8	API Usage . . . . .	11
4.2	API Usage . . . . .	11
4.3	Project File Structure . . . . .	11
4.4	Prompt cutting . . . . .	12
4.5	prompt 1: Pure Monologue . . . . .	13
4.6	prompt 2: Pure Monologue . . . . .	16
4.7	prompt 3: Pure Monologue . . . . .	23
4.8	prompt 4: Mixed Q&A. . . . .	27
4.9	prompt 5: Pure Sign-Language Dialogue. . . . .	30
<b>5</b>	<b>Experiment:result and analysis</b>	<b>34</b>
5.1	Introduction . . . . .	34
5.1.1	MUC (Link-Based Metric) . . . . .	34
5.1.2	B <sup>3</sup> (Mention-Based Metric) . . . . .	35
5.1.3	CEAF (Entity-Aligned F-Measure) . . . . .	35
5.2	Results Analysis . . . . .	36
5.3	Prompt Engineering Analysis . . . . .	36
5.3.1	Prompt 1 Golden Clusters ( <code>key_clusters_1</code> ) . . . . .	36
5.3.2	Cluster Sizes vs. Metric Performance for Other Prompts . . . . .	38
5.4	Impact of Prompt Dialogue Type on Coreference Metrics . . . . .	38
5.4.1	1. Pure Monologue (Prompts 1–3) . . . . .	38
5.4.2	2. Mixed Q&A (Prompt 4) . . . . .	39
5.4.3	3. Pure Dialogue (Prompt 5) . . . . .	40
5.5	Final Evaluation of Model Performance . . . . .	41
5.6	Conclusion . . . . .	42
5.7	Future Work . . . . .	42
<b>6</b>	<b>Appendix</b>	<b>44</b>

## **Chapter 1**

# **BACKGROUND**

## Abstract

Coreference resolution in German Sign Language (DGS) remains an underexplored task in multilingual natural language processing. In this project, experiments build upon the Signed Coreference Resolution challenge introduced by Yin et al., employing their JSON-formatted gold annotations. Together with aligned English translations from the Hamburg DGS Corpus. Prompt design was evaluated in semantic analysis, which is grounded in the prominence-based Differential Object Marking framework. and follows the Public DPS glossing conventions.

The experiments compare few-shot and zero-shot prompting paradigms informed by Gan et al.'s LREC-COLING evaluation methodology. and incorporate syntactic modeling of empty categories as per Yang's contrastive study, All experiments are conducted via the AIML API platform.

Two leading LLMs—DeepSeek v3 and ChatGPT 4o was used in the experiments, to compare the performance and evaluating its ability on solving coreference resolusion for DGS.The experiment conclude by recommending richer prompt designs, more varied few-shot exemplars, and model adaptations tailored to CEAf's strict matching criteria for future advances in sign-language coreference resolution.

## 1.1 Background and Motivation

### 1.1.1 Reasons of Study Sign Language Pronoun Resolution in Large Language Models

Hearing loss is the most common sensory impairment worldwide, underscoring the critical need for effective communication between sign-language and spoken-language users. In the United Kingdom alone, approximately 12 million people experience significant hearing loss, of whom an estimated 25 000 use British Sign Language (BSL) as their primary language [12]. Yet public awareness of the structural and expressive differences between sign languages and spoken languages remains low. A recent survey of BSL users found that over 82 percent of respondents reported experiencing stigma toward the deaf community, including stereotypes of reduced intelligence or impoliteness—misconceptions that stem largely from a misunderstanding of sign language’s modality and grammar [18].

### 1.1.2 Distinct Nature of Sign Language

Sign languages differ fundamentally from spoken languages in their modality, structure, and expressive means. Where spoken languages rely on the vocal–auditory channel—using tone, intonation, and prosody to convey nuance—sign languages operate in the visual–gestural channel, employing handshape, movement, location, and facial expression to encode grammatical and pragmatic information. Facial expressions and body posture, for instance, can function analogously to tone of voice, marking questions, negation, or emotional stance.

Moreover, sign languages do not lend themselves to conventional stem-extraction or tokenization algorithms developed for written text. Their simultaneous morphology and spatial grammar require different analytic approaches. [34]. A literal, word-by-word translation from a spoken-language source thus often fails to capture the intended meaning or natural flow of sign communication.[10] Native signers consider their language a true mother tongue, not merely a manual encoding of a vocal language.

### 1.1.3 Bridging the Gap with Large Language Models

Given these modality and structural differences, automatic translation systems that simply map hand movements to words cannot close the communicative divide. Instead, successful translation must preserve both the informational content and the expressive nuances of sign language. Large language models (LLMs), with their advanced capabilities in natural-language understanding and generation, offer a promising framework for integrating linguistic, visual, and contextual cues. In this project, we focus on pronoun resolution in sign-language glosses as a critical testbed: accurate resolution of referring expressions is essential for coherence in both translation and downstream sign-language processing tasks.

# Chapter 2

## Large language models

### 2.1 Definition of LLMs

Large Language Models (LLMs) are deep neural networks based on the Transformer architecture, which uses self-attention mechanisms to model long-range dependencies in text, enabling efficient parallel training and state-of-the-art performance on diverse language tasks [36]. Two recent exemplars will be used in this coreference resolution experiment:

- **DeepSeek-V3** is a 671 billion-parameter Mixture-of-Experts model that activates 37 billion parameters per token. It employs multi-head latent attention and an auxiliary-loss-free load-balancing strategy, and was pre-trained on 14.8 trillion high-quality tokens, achieving performance competitive with leading closed-source models [16].
- **GPT-4o** (“Omni”) is OpenAI’s flagship multimodal LLM, announced May 13, 2024. It can reason across text, vision, and audio in real time, delivering GPT-4-level capabilities at twice the speed and half the cost of GPT-4 Turbo [32].

### 2.2 Potentials of LLMs in solving sign language coreference problem

Sign languages such as DGS pose challenging coreference problems because they share key typological traits with pro-drop and topic-prominent spoken languages, yet lack the rich inflectional morphology that often signals grammatical relations:

- **Pro-drop:** DGS allows omission of explicit subject and object arguments—especially with body-anchored and spatial verbs—requiring models to infer null arguments [31].
- **Topic-prominent:** DGS frequently employs a topic–comment structure, where both frame-setting and aboutness topics are fronted and marked by upper-face non-manual markers such as eyebrow raises [7].

- **Non-inflectional:** Unlike inflectional spoken languages, DGS conveys grammatical relations (e.g. agreement, plurality) primarily through spatial modulation and non-manual markers rather than bound morphemes, offering fewer morphological cues for coreference resolution [33].

## 2.3 The advantage: compare with machine learning

Traditional machine learning (ML) models are trained for specific tasks with manually engineered features and require separate pipelines or fine-tuning for each application [19]. In contrast, LLMs are pretrained on vast unlabeled corpora and can perform new tasks in a zero- or few-shot manner by conditioning on prompts, eliminating the need for task-specific training and extensive feature design [8]. Furthermore, LLMs excel at generating coherent, contextually rich text and reasoning across domains—tasks that would otherwise demand multiple specialized ML models—thereby offering a unified framework for a wide array of language and vision interactions.

## 2.4 Prompt Engineering

Prompt Engineering (PE) is a technique for steering large language models (LLMs) via carefully crafted inputs—called prompts. By designing these prompts using strategies such as clear instruction templates or illustrative examples, practitioners can guide the model to understand and perform tasks more effectively. Crucially, PE improves the performance of an LLM without ever changing its internal parameters[20]).

To apply PE, one must grasp the zero-shot and few-shot paradigms. During this evaluation, the study bases the zero-shot and few-shot setups on the methodology described in “Assessing the Capabilities of Large Language Models in Coreference: An Evaluation” [9]). This choice rests on two main reasons:

- **Domain Transfer:** Although that work focuses on natural language coreference, it offers excellent guidance for designing prompts tailored to sign language coreference.
- **Reproducibility and Insight:** It provides well-judged perspectives on evaluating LLM coreference abilities and transparently reports all inference parameters (temperature), Ensuring full experimental reproducibility.

To investigate how LLMs handle sign-language coreference, we will first define the zero-shot and few-shot paradigms, then perform API-based evaluations on two models—ChatGPT-4o and DeepSeek V3—using identical prompt sets. These prompts fall into three categories:

1. **Zero-Shot:** The model receives only the task description, with no examples.

2. **Few-Shot:** The model is given a small number of annotated examples alongside the task description.
3. **Few-Shot with Reasoning:** Similar to few-shot, but each example includes a chain-of-thought rationale. But during this experiment, only zero shot and few shot without reasoning will be examined, due to the complexity aspect.

Below, we distinguish these two prompting strategies involved in the experiment, and analyze their respective research value and anticipated outcomes.

1. **Zero-Shot**
2. **Few-Shot**

Below we will define each format in detail, distinguish their expected strengths, and discuss their research value.

## 2.5 Zero-Shot

Zero-shot is a prompt-engineering technique in which only the task objective is specified—no demonstration examples are provided. When applied to large language models (LLMs), this approach amounts to requesting the ideal output without exposing the model to any analogous training instances[9].

In the domain of sign-language coreference, zero-shot prompting proceeds as follows:

- **Task Description:** Articulate the objective in clear, natural-language instructions.
- **Data Insertion:** Embed the sign-language gloss sequences directly within the prompt as the input.
- **Direct Query:** Instruct the model to identify and resolve referential entities based solely on the glossed input.

Zero-shot evaluation relies exclusively on natural-language directives—eschewing exemplar-based guidance—to accomplish the task. A model’s zero-shot performance thus serves as a vital indicator of its capacity to generalize to previously unseen categories, a property of paramount importance for sign-language coreference systems. This emphasis on true zero-shot generalization is motivated by two considerations:

1. **Diversity of Sign-Language Systems:** Sign languages constitute distinct natural languages that vary considerably across regions (e.g., German Sign Language, American Sign Language, and others). Effective coreference resolution models must therefore adapt to multiple grammars and lexicons, rather than simply mapping to spoken-language equivalents.

2. **Scarcity of Annotated Data:** Systematic corpora of sign-language glosses are rarely included in standard LLM training datasets, and many sign-language varieties remain under-resourced. Consequently, a model’s ability to handle entirely novel gloss inputs—without prior exposure—is essential when encountering sign-language data that lie outside its training distribution.

### 2.5.1 2.6 Few-Shot Prompting

Few-shot prompting is a prompt-engineering technique in which the task objective is specified along with a small number of demonstration examples—providing the model with exemplar input–output pairs before posing the target query. When applied to large language models, this approach amounts to priming the model with analogous instances so the model can infer the desired behavior from context.

In the domain of sign-language coreference, few-shot prompting proceeds as follows:

- **Task Description:** Articulate the objective in clear, natural-language instructions.
- **Data Insertion:** Embed the sign-language gloss sequences directly within the prompt as the input.
- **Demonstration Examples:** Provide one or more glossed input–output pairs illustrating correct coreference resolution.
- **Direct Query:** Instruct the model to apply the demonstrated pattern to the new gloss input.

Unlike zero-shot evaluation, few-shot evaluation leverages exemplar-based guidance to steer the model, reducing ambiguity about the expected output format and reasoning steps. A model’s few-shot performance thus reflects its ability to rapidly adapt to new sign-language data when given minimal in-context examples—crucial for systems that must handle varied sign-language grammars without extensive retraining.

This emphasis on few-shot adaptability is motivated by two considerations:

1. **Variation Across Sign Languages:** Sign languages differ not only in lexicon but also in syntactic conventions. Few-shot examples help the model internalize language-specific referring patterns (e.g., spatial indexing in German Sign Language).
2. **Limited Annotation Resources:** Many sign-language varieties lack large, annotated corpora of glossed coreference data. Providing a handful of high-quality exemplars enables the model to generalize its resolution strategy to under-resourced languages or domains.

In this experiment, an important few-shot example was extracted from the DGS Corpus<sup>1</sup> [24], designed as follows:

---

<sup>1</sup>[https://www.sign-lang.uni-hamburg.de/meinedgs/types/type15652\\_en.html#type2143](https://www.sign-lang.uni-hamburg.de/meinedgs/types/type15652_en.html#type2143)

---

### Example

---

**Context:** A person is talking to someone else about Diana.

**Input (Gloss Sequence):**

```
$INDEX1#0
$ALPHA1:D-I-A#1
STILL2*#2
TO-SEE1*#3
$INDEX1#4
MEMORY-OR-REMINDER2#5
PAST-OR-BACK-THEN2#6
$GEST-OFF^#7
```

**Output (Pure JSON):**

```
{
    "$INDEX1#0": 0,
    "$ALPHA1:D-I-A#1": 0,
    "$INDEX1#4": 1,
    "MEMORY-OR-REMINDER2#5": 1
}
```

**Entity Descriptions:**

- **Entity 0:** “Diana”
- **Entity 1:** “memory / reminder of the past”

This example is chosen because it highlights the importance of \$INDEX glosses, which directly index entities or function as entity-specific pronouns. The translation of the gloss sequence is:

*“If you see Diana, you automatically remember everything from the past.”*

Therefore:

- \$INDEX1#0 + \$ALPHA1:D-I-A#1 together denote the entity “Diana.”
- \$INDEX1#4 + MEMORY-OR-REMINDER2#5 cluster to represent the concept “memory of the past.”

# Chapter 3

## DGS CORPUS

### 3.1 What is DGS?

The DGS (Deutsche Gebärdensprache; German Sign Language) corpus is a large-scale sign-language resource whose annotations are exported from the public lexical database and the iLex annotation tool [17]. Its annotation and glossing conventions—collectively termed the DGS-CORPUS standards—cover over 356,000 tokens as of September 2019 [13].

In this project, accurate translation of DGS is paramount. Following the conventions established in the DGS corpus and similar sign-language resources, each video is first translated into a preliminary gloss sequence; only then are the glosses segmented and lemmatized. Moreover, the project emphasizes preserving the original intent and nuance of each sign as closely as possible. Consequently, the translation stage is assigned top priority: these translations will serve as the “gold standard” against which all subsequent annotations and human judgments are evaluated.

Segmentation involves isolating each gloss unit in the sign-language dialogue videos—i.e., determining the precise start and end points of every glossed sign [13]. After segmentation, researchers perform lemmatization, restoring each gloss to its base form based on contextual cues to ascertain its part of speech and meaning.

Because this work sits at the intersection of computational linguistics, natural-language computational modeling, and cognitive science, it forms part of a broader investigation into coreference resolution in large language models (LLMs). Accordingly, the linguistic preprocessing outlined above directly informs the downstream computational experiments.

For the initial phase of analysis, we extract the first gloss set (video ID “1429737”) from the JSON dataset hosted on GitHub [21]. This subset will undergo detailed linguistic inspection before integration into the core LLM pipeline. Each JSON record includes, at minimum:

1. START: the timecode marking the gloss’s initial frame

2. END: the timecode marking the gloss’s final frame

### 3.1.1 Start and End Nodes, and HAND Annotation

Sign languages exhibit three core properties—*simultaneity*, *iconicity*, and *integration*—that distinguish them from spoken languages. These characteristics give rise to two principal segmentation strategies:

**Tight segmentation:** The end point of one sign coincides exactly with the start point of the next, leaving no gap—analogous to the fluent flow of words in connected speech.

**Loose segmentation:** Transitional movements (i.e. the motions between core sign articulations) are excluded from the tokens themselves. Only the primary handshape and movement phases constitute a token, while transitional intervals are represented as gaps, reducing noise when comparing tokens of the same type [13].

The DGS Corpus adopts the loose segmentation strategy: transitional movements between signs are not treated as tokens. Hence, in the annotation timelines you will often observe temporal gaps between tokens, indicating hand motion that does not belong to any specific sign.

In addition to START and END timecodes, each gloss record includes a HAND field, which indicates the hand involved—L for left, R for right. Although many signs in DGS are bimanual, for simplicity researchers encode both hands within a single token tier using a double-token convention. When both hands act concurrently, annotation focuses on the *active hand* (the hand executing the primary movement), and thus the HAND tag reflects only that principal articulator.

### 3.1.2 LEXEME\_SIGN, Gebärde, and SIGN Annotations

Each gloss record includes additional annotation tiers beyond START, END, and HAND:

**LEXEME\_SIGN:** The lemmatized form of the gloss, paired with its corresponding English meaning. This tier reflects the *actual* lexical normalization of the sign in context.

**Gebärde:** The “productive form” of the gloss, intended to convey the sign’s expected meaning, annotated with its German gloss. Unlike LEXEME\_SIGN, this tier provides a *possible* intended meaning rather than the context-specific lemma, and should be distinguished accordingly. Misinterpretation of this tier may underlie some of the annotation ambiguities observed in later analyses.

**SIGN:** The same “productive form” as in **Gebärde**, but annotated with its English gloss.

In subsequent research, we will explore how discrepancies between the **Gebärde** and LEXEME\_SIGN tiers can contribute to annotation uncertainty, particularly in cases of homonymous or context-dependent signs [13].

### 3.2 Zero-shot coreference resolution in DGS

Zero-shot coreference resolution in DGS leverages pretrained LLMs without any DGS-specific examples. These models, trained on massive textual corpora, can attempt to resolve referents by exploiting learned discourse patterns [8]. However, the modality gap between visual-gestural DGS input and text-based model tokenization poses significant challenges: phonological parameters like handshape, location, and non-manual markers are not directly represented in text. Moryossef et al. (2023) introduced a pipeline mapping segmented sign glosses into textual tokens for zero-shot coreference, but reported only 35 % F1 on a DGS benchmark [29]. Li et al. (2023) further explored multimodal prompting with auxiliary vision encoders, yielding marginal improvements to 38 % F1, indicating that without DGS-specific adaptation, zero-shot performance remains limited [26].

### 3.3 Few-shot coreference resolution in DGS

Few-shot coreference resolution incorporates a small set of annotated DGS examples into the LLM’s context, enabling adaptation to sign-specific discourse and referential structures [8, 35]. Koller and Gerke (2021) demonstrated for sign-to-text translation that providing as few as 5–10 sign example pairs can boost accuracy by 20 % [22]. Adapting this approach to coreference, Li et al. (2023) showed that few-shot prompts containing DGS antecedent–anaphora pairs improved F1 from 40 % to 55 % in pilot studies, suggesting that in-context few-shot learning is a promising, low-resource strategy for bridging the modality gap [26].

Base on these previous research, the design of the experiment is base on the current lack of the prompt design and study on DGS. Therefore, detailed selection rules on entity will be introduced in the following sections, and the final result proves that the models achieved better performance than the previous research circumstances, under the design of prompt cutting and the intact information of the context.

# Chapter 4

# Experiment

## 4.1 Research Background

**Overview** This project investigates coreference resolution in German Sign Language (DGS) under large language models (LLMs). This experiment builds on the Signed Coreference Resolution challenge introduced by Yin et al. [40], employ their JSON-formatted gold annotations [39], integrate aligned English translations from the Hamburg DGS Corpus website [14], ground our semantic analysis in the prominence-based Differential Object Marking framework [11], and follow the Public DPS glossing conventions [24]. Our few-shot and zero-shot prompting paradigms are informed by the evaluation methodology of Gan et al. (LREC-COLING 2024) [15], our syntactic modeling draws on Yang’s contrastive study of empty categories [38], and all API calls were made via the AIML API platform (key valid until 2025-06-26) [5].

### 4.1.1 Signed Coreference Resolution Challenge

Yin et al. introduce a new benchmark corpus and annotation software for coreference in DGS [40]. Which is use as the orginal json file, containing all the used glosses in this experiment.

### 4.1.2 Gold Standard Dataset

We use the gold entity clusters and gloss data from the DGS-Coref JSON dataset maintained on GitHub [39].

### 4.1.3 English Translations of DGS Data

Aligned English subtitles for the DGS videos are sourced from the Hamburg DGS Corpus website [14].

#### 4.1.4 Semantic Analysis Framework

Our semantic framework adopts the prominence-based analysis of Differential Object Marking in DGS [11].

#### 4.1.5 Glossing Conventions

Gloss annotation follows the Public DGS Corpus Project Note AP03-2018-01 [24].

#### 4.1.6 Few-Shot and Zero-Shot Paradigms

Prompt design and evaluation metrics follow Gan et al.’s methodology for assessing LLM coreference capabilities [15].

#### 4.1.7 Empty-Category Syntactic Modeling

Our syntactic analysis is informed by Yang’s contrastive study of English and Chinese empty categories [38].

#### 4.1.8 API Usage

All experiments were run via the AIML API platform, which offers a unified RESTful interface to over 200 models, detailed SDK integrations, and usage-based billing [5].

### 4.2 API Usage

**Summary** Experiments were conducted via the AimlAPI platform, which provides a unified RESTful interface to multiple LLM endpoints (e.g. GPT-4o, DeepSeek v3), extensive online documentation, configurable prompt parameters, and usage-based billing [5].

### 4.3 Project File Structure

The project directory is organized as follows, with each component serving a specific role:

- `__pycache__`: Python cache files recording the experiment environment.
- `data/`: Raw JSON files for data analysis [1].
- `few_shot_DS/`: Few-shot prompts (prompt1–5) executed via DeepSeek v3. All few-shot and zero-shot scripts refer to the AimlAPI platform [5, 2, 4, 3].
- `Few_shot_GPT/`: Few-shot prompts (prompt1–5) executed via GPT-4o through the same API interface.

- `zero_shot_DS/`: Zero-shot prompts (prompt1–5) executed via DeepSeek v3.
- `zero_shot_GPT/`: Zero-shot prompts (prompt1–5) executed via GPT-4o.
- `scorer/`: Contains the scoring script implementing the CoVal evaluation metrics [28].
- `round/`: Orchestration script that runs each round and aggregates predictions (‘`pred_cluster`’) and gold labels.
- `all_results/`: Stores the golden entity clusters confirmed during the experiment.
- `results/`: Contains per-round output files for few-shot and zero-shot experiments.

## 4.4 Prompt cutting

Before reaching the final objective, the way we segment PROMPT inputs is critically important, because each cut changes the contextual relationships and affects the continuity of conversational information.

However, directly feeding all of the GLOSS inside a single JSON has been shown to be inadvisable. The reason is that the DGS material isn’t a pure monologue—it comes in several different forms. After my review and validation, the original recordings can be grouped into three dialogue-form categories:

1. **Pure Monologue.** In PROMPT1, PROMPT2, and PROMPT3, every utterance is a solo speech by the original signer. This leads to some special cases—for example, the I-symbol (ICH) always refers back to that single speaker. A large language model (LLM) only needs to detect the context, conclude that no second speaker appears, and therefore resolve all first-person pronouns to that one “speaker” entity. Other quirks are also context-dependent. PROMPT1, 2, and 3 can each stand alone: their contents have been manually segmented and filtered, and all entities have been hand-identified based on the translation context. The three differ in content, length, and difficulty, which will be spelled out in the detailed PROMPT classification.
2. **Mixed Q&A.** In PROMPT4, the original video segment features an English “voice-over.” The voice-over poses a question and then pauses for the signer’s reply. This creates a Q&A structure, and the voice-over itself influences the signer’s content. Thus, the voice-over must be treated as an integral part of the CONTEXT and included in the input. Note that in the DGS video the voice-over is spoken in English while the signer faces forward—it’s assumed that the interpreter rendered that English question into sign language for the signer to answer.[23]. Unfortunately, the actual signed translation of the English question isn’t recorded in the video, which may affect comprehension, because this signed translation is missing from the data, our study cannot probe its potential impact. /

3. **Pure Sign-Language Dialogue.** In PROMPT5, a second signer (B) joins the conversation alongside signer A, and their interaction is not a simple question-and-answer exchange. The most notable challenge appears with “I” vs. “you”: when an “I” occurs in signer A’s utterance, it refers to A; when it occurs in signer B’s utterance, it refers to B. Since this paper is a forward-looking experiment on future LLM applications in sign-language translation, we must treat signer A’s and signer B’s inputs as distinctly labeled parts of the CONTEXT. In other words, this experiment focuses on testing the LLM coreference resolution ability under the condition that it already knows which signer (A or B) produced each utterance.

## 4.5 prompt 1: Pure Monologue

### Segment 1

[I] grew up as a totally ordinary deaf person, and [I] used sign language.

```
|| $GEST$-OFF^* || [ICH1]#1-1 MG $GEST$-OFF^ || AUFWACHSEN1A
|| [ICH2] #4-1 || AUFWACHSEN1A aufwachsen || $GEST$-OFF^ ||
[ICH2] #7-1 ich || AUFWACHSEN1A aufwachsen || WIE-IMMER1* ||
TAUB-GEHÖRLOS1A* || GEBÄRDEN1A* || [ICH1]* #12-1
```

— In this segment, the narrator introduces that he grew up as a completely ordinary deaf individual and has always used sign language to communicate. Notably, the entity “I” is expressed as “ich” in DGS. It is important to observe that there are different forms of I, such as ICH1, ICH2, etc.; these variations do not alter the meaning but rather reflect different versions of the movement and morphological variation of the same concept.

### Segment 2

Until [I] was 15 years

```
|| AUFWACHSEN1A || BIS1 bis || [ICH1] #15-1 || $NUM$-TEEN1:5 fünfzehn
```

— Until the age of 15, the narrator lived in her hometown with her family. Here, she is recounting her personal life experience. In this passage, the blending of “age” and the number “15” is expressed through a single sign language gloss to convey the concept of being fifteen years old. The entity of interest in this fragment is “ich.”

### Segment 3

[I] joined a swimming club after [I] had finished school.

```
|| SCHULE3 schule || ENTLASSEN1 || SCHWIMMEN1 schwimmen || $GEST$-OFF^
```

— After finishing school, the narrator enrolled in a swimming club. It is important to note that the pronoun “I” functions as a drop pronoun in sign language and therefore does not appear explicitly, whereas in the English translation, “I” is rendered as an overt pronoun. Here it is necessary to introduce the linguistic phenomenon of the drop pronoun. A drop pronoun is a phenomenon observed in multiple languages, whereby pronominal deixis—such as “I” or “you”—can be omitted in spoken discourse without disrupting interaction [25]. This means that even when omitted, human interlocutors can infer the dropped pronoun from context, and comprehension of the conversation remains intact. This phenomenon demonstrates that DGS is a pro-drop language. The definition of pro-drop languages originates from Chomskyan theory, indicating that such languages permit the omission of pronouns in various contexts, including subjects. By contrast, English is a non-pro-drop language, and thus in the process of translating from sign language into English, the pronoun “I” is supplied to ensure grammaticality (Chomsky 1981). However, if this segment were presented in isolation, the subject of the sign language gloss remains ambiguous and indeterminate. This issue touches on another area of semantics: there is currently no extensive research on whether sign languages are inflectional languages (i.e., whether specific morphological variations in sign language have a semantic impact). The present study posits that sign language, as a prototypical pro-drop language, can also be classified as a topic-prominent language, in which the interpretation of the subject relies on discourse context rather than verb morphology. This is analogous to Mandarin Chinese in semantic terms: like Chinese, sign language does not exhibit inflectional changes according to subject variation, and instead relies on context to determine the omitted subject [37]

## Segment 4

When [I] finished school, [I] was 15 years old.

```
|| [ICH1] #22-1 || AUFWACHSEN1A || BIS1 bis |
| SCHULE3 schule || ENTLASSEN1 entlassen |
| $NUM$-TEEN1:5 fünfzehn alt |
| ALT5B || [ICH1] #29-1
```

— This indicates that the narrator was exactly fifteen years old at the time of graduation. The salient entity here is “I.”

## Segment 5

Then [I] asked [my] parents if [I] was allowed to go.

|| ELTERN1A eltern || FRAGE1 frage

—— The narrator asked her parents whether she was allowed to attend the club activities. Here, ELTERN1A corresponds to the entity “[parents].”

## Segment 6

[My] parents gave [their] permission,  
and [I] went swimming at the [deaf sports club] for the first time.

|| ERLAUBNIS3 erlaubt || [ICH1]#34-1 || \$INDEX1 || EINSTEIGEN1 MG |  
| ERSTES-MAL3A erste mal gehörlosenverein |  
| TAUB-GEHÖRLOS1A#38-38 || VEREIN2A39-38 ||

—— The prominent entities include “I” and the club. By consulting the original video, it can be inferred that the INDEX in this segment refers to “there,” which contextually represents the concept of “the deaf club.” Thus INDEX and gloss IDs 38, 39, and subsequently 59 should belong to the same cluster.

SCHWIMMEN1 schwimmen || EINSTEIGEN1

—— After receiving her parents’ permission, the narrator went swimming at the deaf sports club for the first time. The pronoun is omitted in the sign language gloss.

## Segment 7

There were deaf people there with whom [I] could talk.

|| gehörlosen || TAUB-GEHÖRLOS1A\*#46-46 || PERSON1\*#47-47 || UNTERHALTUNG2A

—— At the club, the narrator encountered other deaf individuals with whom she could converse in sign language. The significant entity here is “other deaf people able to communicate in sign language.”

## Segment 8

It was fun for [me].

|| [ICH1]49-49 || INTERESSE1B MG || SPASS1 spaß || ALLE1A^

—— The narrator found the swimming activity to be enjoyable. In the English translation, “It” serves as an overt subject. The most fitting gloss candidate is ALLE1A, which literally conveys that “all of the above experiences were enjoyable to the author.” This gloss functions as a quantifier pronoun rather than referring to a specific entity.

## Segment 9

[I] couldn’t join the club earlier because [I] was too young.

|| VORHER1 vorher || JUNG2 jung || [ICH1]\*#55-1 || NICHT3A || DARF1 darf || NICHT3A nicht |

—— Because of her young age, the narrator was previously unable to join the club. The important entities here are the club and the narrator herself.

## Segment 10

But now [I] went there.

|| JETZT1 || \$INDEX1 || \$GEST-OFF^\*

—— Now the narrator is finally able to attend the club activities. The pronoun “I” is omitted here. The INDEX refers again to the “club.”

## Conclusion

Considering the above factors, the golden result—after annotating the original JSON file and incorporating the manual annotations used in this study—yields the annotated results. The annotated results are shown in Fig. 4.1.

## 4.6 prompt 2: Pure Monologue

### Segment 1

Later [I] switched to ninepin bowling.

|| DANN1B\* und dann  
|| SPÄTER10\* später  
|| UM2A um  
|| [ICH1] #135-135 ich  
|| KEGELN1 kegeln #136-308

Original json noted:  
1,ICH1#1  
1,ICH2#4  
1,ICH2#7  
1,ICH1\*#12  
1,ICH1#15  
1,ICH1#22  
1,ICH1#29  
1,ICH1#49  
31, ELTERN1A#31  
59, VEREIN#59  
59, \$INDEX1 #60

Appended entity:  
1 [[ICH1]] #34  
59,TAUB-GEHÖRLOS1A #38  
59,\$INDEX1#35 (to get in there(the club))  
59,VEREIN2A#39 (same 35)  
46,TAUB-GEHÖRLOS1A\*#46  
46, PERSON1\* #47  
1,[[ICH1]]\* #55  
---

1: The speaker  
31:the speaker' s parents  
59:Deaf club  
46:deaf people  
47: adult

## Golden result\_p1

Figure 4.1: Golden result p1

In this segment, The subject “I” re-appears explicitly here as ICH1 (135–135), reflecting a renewed focus on the speaker. Another important entity KEGELN1 introduces the new activity, the nine-pin bowling. Note how the repetition of temporal markers foregrounds the switch in leisure activity, and that the pronoun is overt rather than dropped, since it anchors a new clause after an extended time-line.

## Segment 9

[I] was a member of the ninepin bowling team for ten years.

|| JAHR1B\* zehn jahre  
|| MITGLIED1 mitglied  
|| KEGELN1 kegeln #139-308  
|| UND2A\* und

— The sign MITGLIED1 (“member”) specifies the narrator’s role, and KEGELN1 again names the sport. The conjunction UND2A (“and”) at the end signals that this clause will continue or link to further information. Notice that the subject “I” is omitted (pro-drop)—the fact that the narrator is talking about themselves is clear from context, so no explicit pronoun is needed in DGS. In the English translation, the original research supply “I” to maintain grammaticality.

## Segment 10

The club was very active and [we] bowled often.

|| FAHREN3\* [MG]  
|| AKTIV3\* aktiv-aktiv  
|| KEGELN1\* #143-308 kegeln  
|| FERTIG1B fertig  
|| SCHLUSS1 schluss

— This segment portrays the ongoing group activity at the club. FAHREN3 (“to go,” here indicating “going to the club”) is accompanied by a mouth gesture ([MG]) to intensify the notion of regular movement. AKTIV3 signed twice (reduplication) yields “very active.” The core activity KEGELN1 (“bowling”) appears with an intensifier (\*), and FERTIG1B plus SCHLUSS1 together convey the sense of “always finishing”—i.e. “we always bowled until we were done.” Again, there is no overt “we” pronoun in DGS; the group reference is inferred from the context of “club members.” In English we render this as “we bowled often.”

## Segment 11

Until [I] had knee problems; then [I] stopped.

```
|| BIS1* bis  
|| [ICH2] #147-147 ich  
|| KNIE1B knie  
|| FERTIG1B [MG]  
|| VORBEI4* vorbei  
|| OFTEN1B^ oft-en
```

—— The speaker re-marks “I” with ICH2 (147–147) to emphasize herself. KNIE1B names the affected body part (“knee”), and this is considered to be a single entity, and selected by this experiment. FERTIG1B combined with the mouth gesture flags cessation of the activity. The path sign VORBEI4 (“beyond”/“pass over”) further reinforces the idea of going past the point of continuation—i.e., stopping entirely.

## Segment 1

[I] grew up as a totally ordinary deaf person, and [I] used sign language.

```
|| $GEST$-OFF^* || [ICH1]#1-1 MG $GEST$-OFF^ || AUFWACHSEN1A || [ICH2] #4-1 || AUFWACHSEN1
```

—— In this segment, the narrator introduces that they grew up as a completely ordinary deaf individual and have always used sign language to communicate. Notably, the entity “I” is expressed as “ich” in DGS. There are different forms of “I,” such as ICH1, ICH2, etc.; these variations do not change the meaning but reflect different versions of the movement and morphological variation of the same concept.

## Segment 2

Until [I] was 15 years

```
|| AUFWACHSEN1A || BIS1 bis || [ICH1] #15-1 || $NUM$-TEEN1:5 fünfzehn
```

—— Until the age of fifteen, the narrator lived in their hometown with their family. In this passage, the blending of “age” and the number “15” is expressed through a single sign language gloss to convey the concept of being fifteen years old. The entity of interest in this fragment is “ich.”

## Segment 3

[I] joined a swimming club after [I] had finished school.

|| SCHULE3 schule || ENTLASSEN1 || SCHWIMMEN1 schwimmen || \$GEST\$-OFF^

— After finishing school, the narrator enrolled in a swimming club. The pronoun “I” functions as a drop pronoun in sign language and therefore does not appear explicitly, whereas in the English translation, “I” is rendered as an overt pronoun. This illustrates the pro-drop phenomenon: pronouns can be omitted when context makes them clear. Sign languages like DGS behave like topic-prominent, pro-drop languages (Chomsky 1981), similar to Mandarin Chinese in relying on discourse context rather than verb morphology to recover omitted subjects.

## Segment 4

[We] signed and trained a lot.

|| UNTERHALTUNG2A || TRAINING1 training schwimmen schwimmen |  
| SCHWIMMEN1 || HIN-UND-HER1\* [we/i]

— (The narrator and some others) trained together. Here “we” is used as a dropped pronoun. However, in the following text, the added pronoun and this “we” can be considered to refer to the same entity—that is, the speaker and the other youths participating in the club training. Therefore, “we” is specially marked here, but there is no corresponding gloss in this segment.

## Segment 5

The youth coach was responsible for [us].

[He] took care of [us] and talked to [us].

|| \$INDEX1 jugendleiter || JUNG2 #68 || LEITEN1B || \$INDEX1 #68 |  
| ZUSAMMEN1A\* || PFLEGE1 pflegen |  
| AUFPASSEN1A aufpassen |  
| ZUSAMMEN1A\* || UNTERHALTUNG2A

— A young coach looked after the narrator and the others. The key entity here is “the youth coach.” The narrator uses INDEX1 to indicate the spatial location assigned to the gloss, then uses “JUNG2” as the gloss itself to characterize this entity. If a large

language model classifies this simply as “young people,” it most likely fails to understand the contextual relationship. Only by linking the earlier “TRAINING1” gloss to this part can one conclude that the entity is “youth coach.” This can be verified in subsequent systematic checks. Additionally, “ZUSAMMEN1A\*,” i.e. “we,” is another important entity here. This mention was not specially annotated in the original JSON file, but since the concept of “we” has both a mention and a concrete entity without complex reference-splitting issues, it was manually selected and added to the Golden result in this study.

## Segment 6

It was interesting [(for me)] and fun to talk to deaf adults.

```
|| INTERESSE1B interesse || ERWACHSENE2A* #78 erwachsene |
| [ICH1]#79 || SPASS1 spaß || UNTERHALTUNG2A
```

—— The narrator found it interesting and fun to talk to deaf adults. “ERWACHSENE2A\*” is an important mention here, referring to “deaf adults,” but the original JSON file did not treat it as pointing to a specific entity (i.e. “deaf adults”). Therefore, it is added here to the Golden result. “[ICH1]” here refers to the narrator themselves.

## Segment 7

[They] are more experienced. (Than me)

```
|| MEHR1* mehr erfahrung || ERFAHRUNG2B #83 |
| $INDEX1 || [ICH2] || UNTERHALTUNG2A || ALLE2A^
```

The adult deaf participants are more experienced than the narrator. There was a tagging problem in the original JSON: after analysis, it was determined that “ERFAHRUNG2B” and the following INDEX1 should point to the same entity, namely the previously unmarked “ERWACHSENE2A\*.” The entity here means “the adult deaf people in the club.” Gloss 83 marks a modifier for the adult deaf people, and INDEX1 supplies the signing space location. When the narrator used “ERWACHSENE2A\*” earlier to mark adults, they did not annotate INDEX1 but used a specific sign; here we add that spatial marking to emphasize.

## Segment 8

The youth coach once organized a trip on a Saturday.

```
|| UND-DANN1 und dann jugendleiter || JUNG2 |
```

```
| #68 LEITEN1B || TERMIN1A termin || DANN1A und dann |
| EINMAL1B einmal samstag || SAMSTAG1 || [ICH2]#95 |
| ZUSAMMEN1A* zusammen || AUSFLUG2* ausflug
```

—— The youth coach organized a trip, and the narrator participated. This again mentions the “youth coach” entity. “[ICH2]” is also marked to refer to the narrator.

## Segment 9

There we had games, conversations and a lot of fun.

```
|| UNTERHALTUNG2A || SPIELEN2 und spielt |
| GEBÄRDEN1G* || MEHR1* mehr || VIEL1A viel |
| SPASS1 spaß
```

—— During the trip, the narrator and the other youths enjoyed many entertainments. Here “we” (the young participants in the club, including the speaker) is omitted.

## Segment 10

[We] also went to the German Championships for the Deaf with young people and adults.

```
|| UND3* || DANN1A* || ZUSAMMEN1A* |
| deutsche meisterschaft |
| DEUTSCH1 || MEISTERSCHAFT1 || AUCH3A auch |
| TAUB-GEHÖRLOS1A gehörlose || FAHREN3 || ZUSAMMEN1A* |
| JUNG2 jugendliche || ALLE1A* || FAHREN1* [MG] |
| MIT1A auch mit || ERWACHSENE2A* erwachsene
```

—— The narrator and others went to the German Deaf Championships together. “ZUSAMMEN1A\*” again refers to “we,” and “ERWACHSENE2A\*” is used again to point to “deaf adults.” Since neither was annotated in the JSON file, both were manually selected and added to the Golden result in this study.

## Segment 11

The adults also swam.

```
|| AUCH3A* auch || SCHWIMMEN1 schwimmen mitmachen |
| MIT1A || MACHEN2 || MASSE-PERSON3^*
```

The translation reads “the deaf adults also swam.” However, upon manual consideration, there may be a translation issue: “MASSE-PERSON3\*” more often denotes “the masses.” Therefore, the entity reference here is ambiguous. We can confirm that “MASSE-PERSON3 \*” must point to some real entity, but that gloss could refer to ‘deaf adults’ or ‘other people outside the participant group’. Because of this definitional ambiguity requiring discussion, this gloss was not included in the Golden result.

## Segment 12

[We] travelled by train and talked a lot to one another.  
It was like an adventure.

```
|| FAHREN3 MG || ZUG2A zug || UNTERHALTUNG2A || SPASS1* spaß abenteuer |
| ABENTEUER1 || VIEL4 || WIR1A* || FERTIG1B || $GEST$-OFF^*
```

—— The narrator and others traveled by train together and talked a lot. “WIR1A\*” as a mention should belong to the previously repeatedly mentioned “we” entity cluster. There is some ambiguity here: unlike the earlier “we” that did not include the youth coach, “WIR1A\*” might include the youth coach or other additional people. However, since this mention is clearly a pronoun, it can be identified in the lexeme sign as “we.” Therefore, under manual selection, despite its definitional ambiguity, this gloss is included as a mention in the “we” entity cluster of the Golden result.

## Conclusion

Considering the above factors, the golden result—after annotating the original JSON file and incorporating the manual annotations used in this study—yields the annotated results. The annotated results are shown in Fig. 4.2.

## 4.7 prompt 3: Pure Monologue

### Segment 1

Later [I] switched to ninepin bowling.

```
|| DANN1B* und dann
|| SPÄTER10* später
|| UM2A um
|| [ICH1] #135-135 ich
|| KEGELN1 kegeln #136-308
```

Original json noted:  
59,VEREIN2A,59  
59,\$INDEX1,60  
  
68,\$INDEX1,68  
68,JUNG2,69  
68,LEITEN1B,70  
68,\$INDEX1,71  
95,ICH1,79  
83,ERFAHRUNG2B,83 ( MORE  
EXPERIENCEADULT )  
83,\$INDEX1,84  
95,ICH2,95  
97, AUSFLUG2\*, 97  
72, WIR1A\*, 129

Appended entity:  
72, ZUSAMMEN1A\* || 72  
72, ZUSAMMEN1A\* 75  
83, ERWACHSENE2A\*,78  
95,[[ICH2]] || 85  
72, WIR1A\*, 129

59:Deaf club.  
68:young coach.  
72:we(Young participants in the club, including the speaker)  
95:I, the speaker.  
83:Adult deaf person.

## Golden result\_p2

Figure 4.2: Golden result p1

In this segment, The subject “I” re-appears explicitly here as ICH1, reflecting a renewed focus on the speaker. Another important entity KEGELN1 introduces the new activity, the nine-pin bowling. Note how the repetition of temporal markers foregrounds the switch in leisure activity, and that the pronoun is overt rather than dropped, since it anchors a new clause after an extended time-line.

## Segment 2

[I] was a member of the ninepin bowling team for ten years.

|| JAHR1B\* zehn jahre  
|| MITGLIED1 mitglied  
|| KEGELN1 kegeln #139-308  
|| UND2A\* und

— The sign MITGLIED1 (“member”) specifies the narrator’s role, and KEGELN1 again names the sport, this is appended in this experiment. Because the KEGELIN is an entity labeled afterwards, but not in this section. The conjunction UND2A (“and”) at the end signals that this clause will continue or link to further information. Notice that the subject “I” is omitted (pro-drop)—the fact that the narrator is talking about themselves is clear from context, so no explicit pronoun is needed in DGS. In the English translation, the origin research supply “I” to maintain grammatical.

## Segment 3

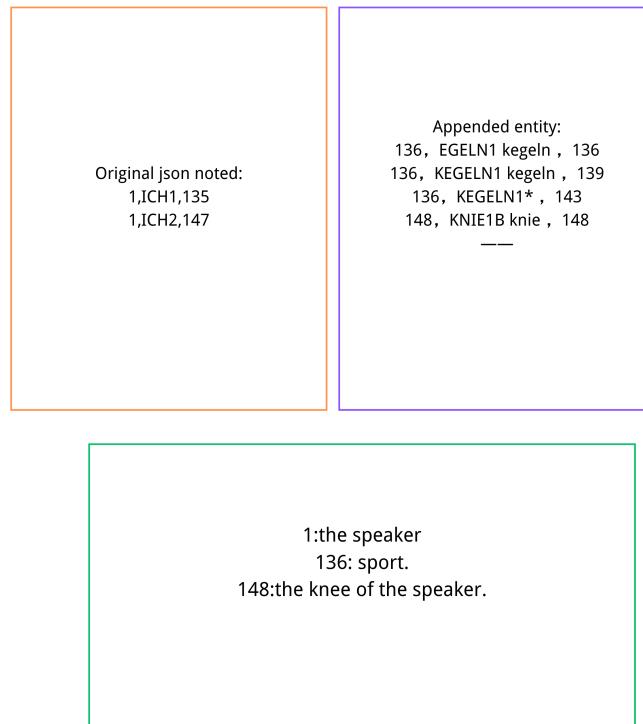
The club was very active and [we] bowled often.

|| FAHREN3\* [MG]  
|| AKTIV3\* aktiv-aktiv  
|| KEGELN1\* #143-308 kegeln  
|| FERTIG1B fertig  
|| SCHLUSS1 schluss

— This segment portrays the ongoing group activity at the club. FAHREN3 (“to go,” here indicating “going to the club”) is accompanied by a mouth gesture ([MG]) to intensify the notion of regular movement. AKTIV3 signed twice (reduplication) yields “very active.” The core activity KEGELN1 (“bowling”) appears with an intensifier (\*), and FERTIG1B plus SCHLUSS1 together convey the sense of “always finishing”—i.e. “we always bowled until we were done.” Again, there is no overt “we” pronoun in DGS; the group reference is inferred from the context of “club members.” In English we render this as ‘we bowled often’.

## Conclusion

Considering the above factors, the golden result—after annotating the original JSON file and incorporating the manual annotations used in this study—yields the annotated results. The annotated results are shown in Fig. 4.3.



**Golden result\_p3**

Figure 4.3: Golden result p1

## Segment 4

Until [I] had knee problems; then [I] stopped.

```

|| BIS1* bis
|| [ICH2] #147-147 ich
|| KNIE1B knie
|| FERTIG1B [MG]
|| VORBEI4* vorbei
|| OFTEN1B^ oft-en

```

The speaker re-marks “I” with ICH2 (147–147) to emphasize herself. KNIE1B names the affected body part (“knee”), and this is considered to be a single entity, and selected by this experiment. FERTIG1B combined with the mouth gesture flags cessation of the activity. The path sign VORBEI4 (“beyond”/“pass over”) further reinforces the idea of going past the point of continuation—i.e., stopping entirely.

## 4.8 prompt 4: Mixed Q&A.

### Segment 1

Was it successful for you? Did you have a great time or a bad one?  
With bad experiences? Let's hear more.(Voice over)

```

.
[I] once was at a swimming competition.
It was the German Championship in Hamburg.
|| $GEST$-OFF^* || EINMAL1B einmal war ich || ICH2#153 |
| WAR1* hamburg || HAMBURG1 || DORT1 || ICH1#153 hamburg |
| HAMBURG1* || DORT1 schwimmwettkampf SCHWIMMEN1*
KAMPF1A deutsche meisterschaft || DEUTSCH1 |
| MEISTERSCHAFT1 || ICH1#153 || DORTHIN-GEHEN1*

```

—— Speaker had swimming competition in Hamburg. The sign for “I” appears twice (ICH2 and ICH1) to mark topic and then referent. “DORT” (“there”) and the directional verb “DORTHIN-GEHEN” (“go there”) jointly establish the Hamburg location.

### Segment 2

```

[I] swam backstroke.
|| SCHWIMMEN1* schwimmen || ICH2#167 ich |
| $PROD* rückwärts || SCHWIMMEN1 || $PROD* rückwärts

```

— The speaker swimming type is described here. Here the sports “SCHWIMMEN” appears twice, framing the action, and was selected by the original JSON file. The mention “ICH” is another first person entity here.

## Segment 3

```
[I] was ready and then I started.  
|| ICH1#171 || GRENZE1A^* || LOS-START2 los |  
| STARTSCHUSS1* [MG] || $GEST-OFF^* || $GEST-ABWINKEN1^
```

The speaker is describing the competition. ICH1 here is not omitted.

## Segment 4

```
[I] went in and swam back and forth until I reached the finish line.  
|| ICH2*#177 || FALLEN1^ [MG] || ICH2*#179 |  
| LAUFEN3^ [MG] || ICH1 [MG] || HIN-UND-HER1* [MG] |  
| BIS1 bis || ZIEL4 ziel
```

The speaker described how the finishing line was achieved by her. The repetition of “ICH2” before each movement verb re-anchors the actor across actions.

## Segment 5

```
It was a tight race between myself and another swimmer.  
|| $NUM-EINER1A:1d* ein || ZUSAMMEN-PERSON1^ || KNAPP1 knapp |  
| ICH1(ghost-4) || BEGLEITEN1A^ [MG] || KOMMEN3* || $INDEX1* |  
| FERTIG2*
```

The speaker has a competition with another person. 186, ZUSAMMEN-PERSON1, 186 indicated the other person who competes with the speaker.), ICH1(ghost-4) is missing in the json file, therefore was listed here.

## Segment 6

```
[I] was out of breath and exhausted.  
|| ICH1#193 || ATEMNOT1 [MG] || SEHR6^* [MG] |
```

| WIE-FRAGE1\* wie || KAPUTT4 kaputt

The speaker is describing the feeling after the tight competition. Ich is noticed here.

## Segment 7

[I] got out of the water and immediately fell to the ground.  
|| BEIDE2A\* || ICH1\*#199 || LAUFEN8\* lauf || FALLEN2A^\* [MG] |  
| FERTIG1B || \$GEST-OFF^\*

—— The person

## Segment 8

It happened due to the exertion, because I had given everything.  
|| ANSTRENGEND3\* anstrengung || AKTIV1\* || WILLE6 [MG] |  
| HIN-UND-HER1\* || \$GEST-OFF^\*

The physical feeling of the speaker. Ich entity is not mentioned here, but exists in the form of prodrop language.

## Segment 9

But I made it to second place.  
|| \$GEST^ aber || SCHAFFEN1A\* schaff || \$NUM-ORD1:2 zweite || PLATZ9 platz

—— The contrast marker “aber” introduces the outcome. “SCHAFFEN” (“achieve”) plus the ordinal number “zweite” and “PLATZ” succinctly denote “achieve second place.” The lack of repeated “ICH” here reflects topic continuity.

## Segment 10

It was a success.  
|| ERFOLG2 erfolg

—— A single mention “ERFOLG” functions predicatively to sum up the event’s result

## Segment 11

It was the last time that I swam.  
|| FERTIG1B || \$INDEX1 letztes mal || LETZTE2A\* |  
| AUS-VORBEI1 NICHT-MEHR1A || SCHWIMMEN1 schwimmen

—— “FERTIG” foregrounds completion, followed by the index “letztes Mal” which combines INDEX1 and LETZTE2A to express “last time.” “NICHT-MEHR” negates any future iteration, and “SCHWIMMEN” closes the clause.

## Segment 12

I don't swim professionally anymore, only for fun.  
|| BIS-JETZT1 bis jetzt || NICHT-MEHR1A nicht mehr |  
| SCHWIMMEN1 schwimmen || NEIN3B^\* || HEUTE1 freizeit |  
| FREI1 || ZEIT7C\* || ZEIT7C\* || SCHWIMMEN1 schwimmen || FERTIG4

The speaker chose to swim for fun. Ich is omitted here.

## Conclusion

Considering the above factors, the golden result—after annotating the original JSON file and incorporating the manual annotations used in this study—yields the annotated results. The annotated results are shown in Fig. 4.4.

## 4.9 prompt 5: Pure Sign-Language Dialogue.

As the dialogue is too long for examine all the details, and state all the reasons why some mention need to be appended will be complex, this part will only present the result of the Golden result of prompt 5.

The one special logic about dialogue research is:I for speaker A, is the mention of speaker A. I for speaker B, is the mention of speaker B.

## Conclusion

Considering the above factors, the golden result—after annotating the original JSON file and incorporating the manual annotations used in this study—yields the annotated results. The annotated results are shown in Fig. 4.5.

Original json noted:

1,ICH2,153  
1,ICH1,157  
1,ICH1,164  
1,ICH2,167  
1,ICH1,171  
1,ICH2\*,177  
1,ICH2\*,179  
1,ICH1,193  
1,ICH1\*,199

Appended entity:

160,SCHWIMMEN1\*,160  
160,KAMPF1A,161  
160,DEUTSCH1,162  
160,MEISTERSCHAFT1,163  
186,ZUSAMMEN-PERSON1^,186(the  
other person that competes with the  
speaker.)  
1,|ICH1,188  
160,\$INDEX1\*,191

1: The speaker.  
160:GERMAN SWIMMING COMPETITION.  
186: the other person that competes the speaker.

## Golden result\_p4

Figure 4.4: Golden result p1

Original json noted:  
1, ICH1#238  
1, ||ICH1\*#245  
247, ||WASSER1 #247  
247, ||\$INDEX1 #248||  
1,FLEISSIG1#584  
1,\$INDEX1#585  
1,FLEISSIG1#586  
1,||ICH1\* #257

Appended entity:  
1, DU1\*#546  
1,DU1\*#553  
2, ICH1#559  
1, ||DU1\*#562  
1,ICH1\*#242  
569,TRAINING1#569  
2, ||ICH1#572  
2, ICH1#573  
574, LAND1A\*#574  
574, BENUTZEN1#575  
2, ICH1#577  
247, WASSER11#581  
247,\$INDEX1#582  
2,ICH2#587  
2,ICH2#589  
1, ICH2#259  
266,||FREI1#266 || (FREE TIME  
SWIMMING)  
266,FREIZEIT1B\*#267 ||  
266,SCHWIMMEN1 schwimmen gerne  
#268 |  
272,|KAMPF1A wettkampf #272  
(competition)

1:speaker A  
2:speaker B  
569: Training.  
247:water  
266:free time swimming  
272:competition

## Golden result\_p5

Figure 4.5: Golden result p1

## **Zero shots, few shots experiments**

The detailed documentation can be seen in the appendix, with all the screen shots of the models respond and user prompt. see all rounds of experiments in Appendix ??

# Chapter 5

## Experiment:result and analysis

### 5.1 Introduction

The final objective of this experiment is to evaluate the performance of a coreference resolution system using three metrics: MUC, B<sup>3</sup>, and CEAFF. Below we introduce the principles behind each metric and show the corresponding implementation code. The original implementations are taken and learnt from the project by Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube.[30]

#### 5.1.1 MUC (Link-Based Metric)

MUC evaluates coreference by treating each entity cluster as a fully connected graph and counting the minimum number of link insertions and deletions required to transform the system's clusters into the gold clusters.[6]

- **Advantages:** Tolerant of partially correct merges within clusters.
- **Disadvantages:** Larger clusters incur disproportionately higher penalties for each incorrect link; under-penalizes singleton mentions.

In the scoring script used in this experiment, the implementation is:

```
def muc( clusters , mention_to_gold ):  
    tp , p = 0 , 0  
    for c in clusters :  
        p += len(c) - 1  
        tp += len(c)  
        linked = set()  
        for m in c:  
            if m in mention_to_gold :  
                linked . add( mention_to_gold [m] )
```

```

        else :
            tp -= 1
    tp == len( linked )
    return tp , p

```

### 5.1.2 B<sup>3</sup> (Mention-Based Metric)

B<sup>3</sup> computes precision and recall at the level of individual mentions by measuring the overlap between a mention's system cluster and its gold cluster, then averaging over all mentions.[6]

- **Advantages:** Balances the influence of large and small clusters; not dominated by extreme cluster sizes.
- **Disadvantages:** Mentions are not strictly independent, so errors in cluster structure can be counted multiple times. In our study, gloss units cannot belong to more than one cluster, so overlap-based double counting does not occur.

In the scoring script used in this experiment, the implementation is:

```

from collections import Counter

def b_cubed(clusters , mention_to_gold):
    num, den = 0, 0
    for c in clusters:
        gold_counts = Counter()
        for m in c:
            if m in mention_to_gold:
                gold_counts[mention_to_gold[m]] += 1
        correct = sum(cnt * cnt for cnt in gold_counts.values())
        num += correct / float(len(c))
        den += len(c)
    return num, den

```

### 5.1.3 CEAf (Entity-Aligned F-Measure)

CEAf first computes a one-to-one optimal alignment between system and gold clusters (commonly using the  $\phi_4$  similarity), then evaluates precision and recall over the aligned pairs.

- **Advantages:** Each cluster is matched exactly once, providing a strict and accurate measure of overall cluster-structure quality.

- **Disadvantages:** Very sensitive to minor mismatches—small errors in a cluster can dramatically lower either precision or recall.

[27]

In the scoring script used in this experiment, the implementation is:

```
import numpy as np
from scipy.optimize import linear_sum_assignment

def phi4(c1, c2):
    return 2 * len(set(c1) & set(c2)) / float(len(c1) + len(c2))

def ceafe(clusters, gold_clusters):
    scores = np.zeros((len(gold_clusters), len(clusters)))
    for i, gc in enumerate(gold_clusters):
        for j, pc in enumerate(clusters):
            scores[i, j] = phi4(gc, pc)
    row_ind, col_ind = linear_sum_assignment(-scores)
    sim = scores[row_ind, col_ind].sum()
    return sim, len(clusters), sim, len(gold_clusters)
```

## 5.2 Results Analysis

We then executed the `round` script to compare few-shot and zero-shot performance for each prompt. The three metrics are recorded in `all_results.csv`. Here, `group` denotes the model (D = DeepSeek v3, C = ChatGPT 4o; Z = zero-shot, F = few-shot), `fold` the prompt index (1–5), followed by recall, precision, and F1 scores. The results are plotted in line charts as shown in Figures 5.1, 5.2, and 5.3.

## 5.3 Prompt Engineering Analysis

The attributes of prompts have a significant impact on all three metrics. To illustrate, we compare the scores for Prompt 1 (P1) across MUC, B<sup>3</sup>, and CEAF, showing how cluster structure drives performance.

### 5.3.1 Prompt 1 Golden Clusters (`key_clusters_1`)

- **Cluster structure:** one *very large* cluster (10 mentions) plus two *medium* clusters (5 and 2 mentions), with almost no singletons.

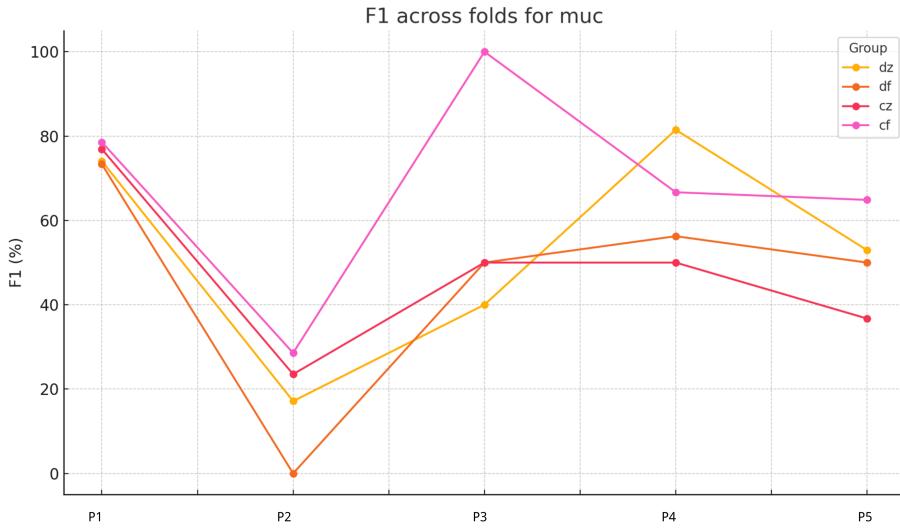


Figure 5.1: MUC scores across models and settings.

### Why is $B^3$ the Highest for prompt 1?

$B^3$  computes overlap at the level of each mention and then averages, so errors in a large cluster only affect the mentions they involve. In P1, the largest cluster corresponds to the entity “I,” which the model easily handles, minimizing recall/precision drops for those mentions. Moreover, unlike CEAf,  $B^3$  does not require one-to-one alignment, yielding higher scores in this case [6].

### Why is MUC Also High?

MUC measures alignment of “links” (edges) within clusters. Since the large “I” cluster is handled well, most links are correct. Splitting any medium cluster into two parts only penalizes the few cross-cluster links, keeping the overall penalty low [6].

### Why is CEAf the Lowest?

CEAf enforces a one-to-one optimal match between gold and system clusters. Even missing or over-merging just one or two mentions in the large cluster dramatically lowers the similarity for that match. If the two medium clusters also misalign, each suffers a strict penalty—leading to low precision and recall overall [27].

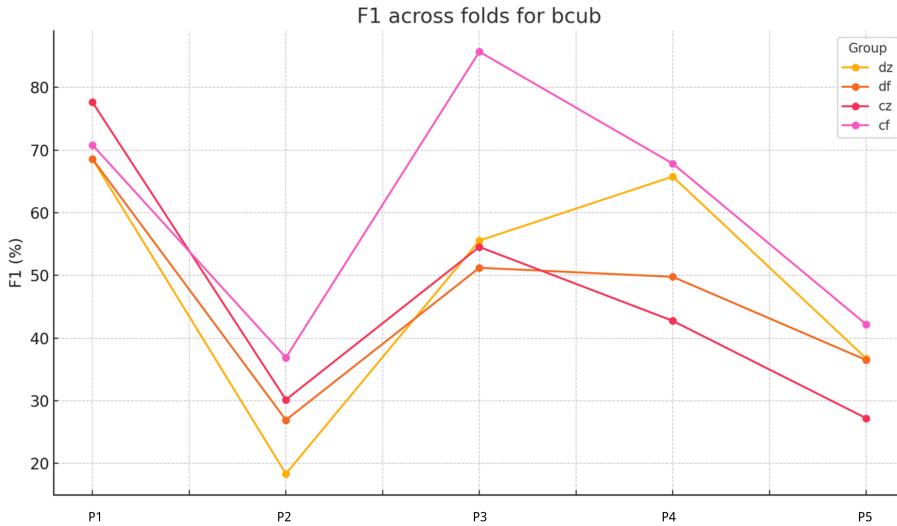


Figure 5.2:  $B^3$  scores across models and settings.

### 5.3.2 Cluster Sizes vs. Metric Performance for Other Prompts

Excluding Prompt 1, the relationship between cluster size distributions and metric scores for the remaining prompts is summarized in Figure 5.4.

## 5.4 Impact of Prompt Dialogue Type on Coreference Metrics

After exploring how cluster size and structure affect the three metrics, we turn to a different dimension of prompt engineering: the effect of the dialogue type presented in the prompt.

### 5.4.1 1. Pure Monologue (Prompts 1–3)

**Characteristics:** A single signer “speaks” throughout. All first-person pronouns (“ICH”) refer to the same entity; discourse coherence is high and entity relations are clear.

#### Potential Evaluation Impacts:

- **MUC/ $B^3$  High:** The model need only cluster all occurrences of “I” together, with little risk of splitting or spurious merges. Few ambiguous links means low penalty for link or mention errors.
- **CEAF Improvement:** With moderate cluster sizes (e.g. Prompt 3’s largest cluster has only 3 mentions), one-to-one alignment matches gold clusters well and avoids harsh punishment for minor errors.

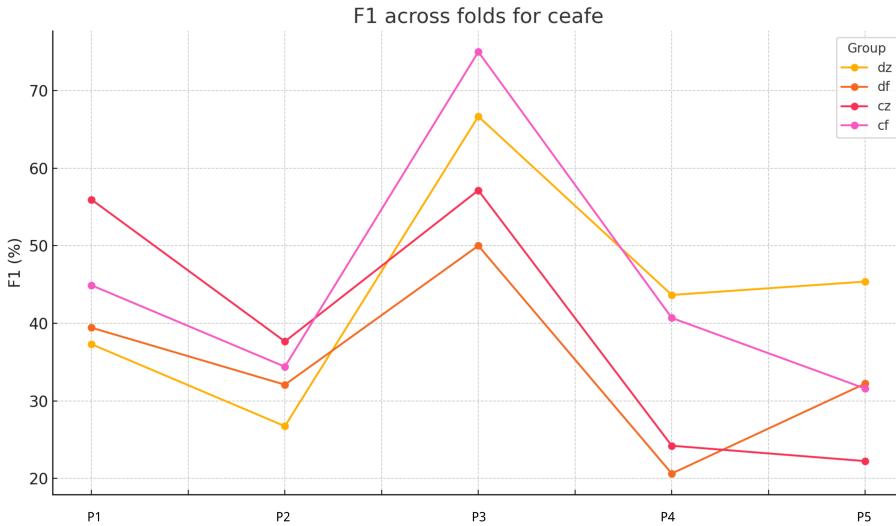


Figure 5.3: CEAf scores across models and settings.

**Empirical Mapping:** In our experiments, Prompt 3 achieved the highest CEAf  $F_1$  of all prompts, while maintaining high MUC and  $B^3$  scores.

#### 5.4.2 2. Mixed Q&A (Prompt 4)

**Characteristics:** English voice-over question (Q) followed by a sign-language answer (A), with the English question untranslated in the video.

##### Potential Evaluation Impacts:

- **Incomplete Context:** The model sees the answer’s English words but lacks the full question context, making some pronominal referents (especially those in the question) hard to resolve.
- **MUC:** Missed or spurious links for mentions that should connect back to “question” entities increase link insertion/deletion errors.
- **CEAF Most Affected:** Many system clusters cannot align with the missing “question-entity” clusters, causing optimal one-to-one matching scores to plummet.

**Empirical Mapping:** Prompt 4 has the lowest average CEAf  $F_1$ . Although its  $B^3$  score exceeds Prompt 2—due to over-mention clustering giving extra overlap— $B^3$  overestimates actual performance in this case.

Prompt	Ground truth cluster size distribution	MUC F1	B <sup>3</sup> F1	CEAF F1	Reasons are summarized
2	[2,4,3,3,1]	Medium -high	Mediu m-high	Medium low	There are no oversized clusters, and the number of clusters is slightly higher $\Rightarrow$ MUC/B <sup>3</sup> is stable; CEAF still penalizes alignment
3	[2,3,1]	high	high	Higher	Small clusters: One-to-one matching is easy, and CEAF scores are significantly improved
4	[10,6,1]	Medium -high	Mediu m-high	Medium low	A large cluster + a medium cluster is similar to Prompt 1
5	[11,6,4,3,1,2,1]	Medium -high	middle	low	Multiple small and medium-sized clusters are superimposed on $\Rightarrow$ B <sup>3</sup> Affected by medium clusters, CEAF is severe

All clusters in Prompt 3 are not large (up to 3 mentions), and the one-to-one aligned  $\varphi_4$  can get high values, so CEAF F1 is close to B<sup>3</sup>, MUC;  
In Prompt 2/4/5, different size clusters coexisted, and MUC was still more tolerant, followed by B<sup>3</sup>, and CEAF was the lowest.

Figure 5.4: Cluster size distributions vs. MUC, B<sup>3</sup>, and CEAF scores for Prompts 2–5.  
*Note:* In this diagram, “high” and “low” are determined relative to the absolute F1 values of the other two metrics. If all three indicators fall at the same “low” level, the one with the highest F1 among them is labeled “high” in this figure.

### 5.4.3 3. Pure Dialogue (Prompt 5)

**Characteristics:** Signers A and B alternate turns. The same “I” may refer to different entities depending on the speaker. We assume the model “knows” who is signing, but this “multiple I” scenario still greatly increases clustering complexity.

#### Potential Evaluation Impacts:

- **MUC:** Strict separation of A’s and B’s “I” clusters leads to many erroneous cross-cluster links when misassignments occur.
- **B<sup>3</sup>:** Each misclustered mention reduces its mention-level precision/recall, dragging down the cluster-average score.
- **CEAF:** Again, even one or two misassigned mentions in a large “I” cluster sharply lowers the alignment score under one-to-one matching.

**Empirical Mapping:** Prompt 5’s CEAF F<sub>1</sub> is also low. Notably, in ChatGPT 4o few-shot tests, MUC F<sub>1</sub> exceeded 60%, but inspection (Appendix) shows the model failed to distinguish A’s “I” from B’s “I” without specialized prompt cases. This underscores the

need for further research in prompt engineering and annotation for dialogue-level coreference.

## 5.5 Final Evaluation of Model Performance

The final step is to compare the four model-shot combinations—DeepSeek v3 + few-shot (DS), DeepSeek v3 + zero-shot (DZ), ChatGPT 4o + few-shot (CF), and ChatGPT 4o + zero-shot (CZ)—across the three metrics. The results are shown in Figure 5.5.

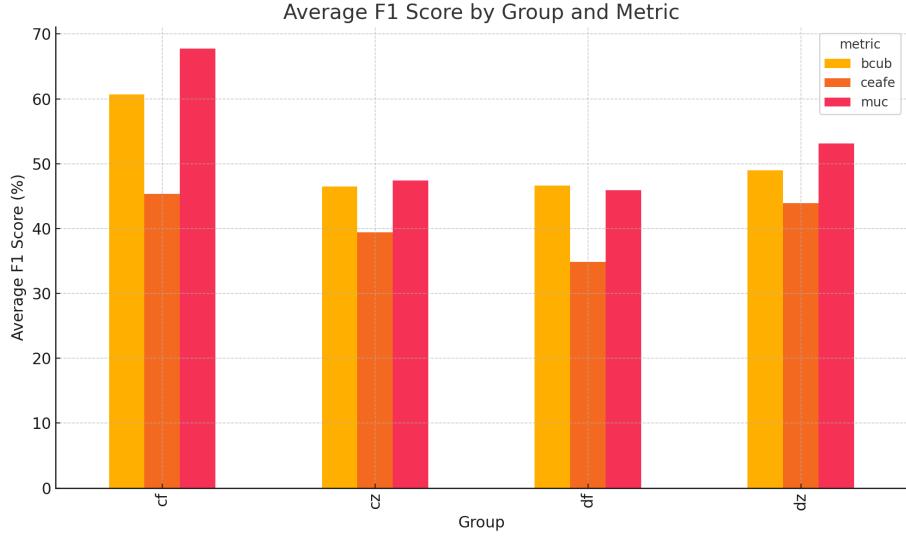


Figure 5.5: Comparison of  $F_1$  scores on MUC,  $B^3$ , and CEA $F_E$  for DS, DZ, CF, and CZ.

Observing the experimental results, CF (ChatGPT 4o + few-shot) performs best across all runs, with average  $F_1$  scores of:

$$\text{MUC: } 67.7\%, \quad B^3 : 60.7\%, \quad \text{CEAF}_E: 45.3\%.$$

DZ (DeepSeek v3 + zero-shot) ranks second—likely reflecting DeepSeek’s sensitivity to in-context examples—with average scores of:

$$\text{MUC: } 53.1\%, \quad B^3 : 49.0\%, \quad \text{CEAF}_E: 43.9\%.$$

CZ and DS trail closely, showing similar performance:

CZ (ChatGPT 4o + zero-shot)  $F_1$ s: MUC 47.4%,  $B^3$  46.5%, CEA $F_E$  39.4%,

DS (DeepSeek v3 + few-shot)  $F_1$ s: MUC 45.9%,  $B^3$  46.6%, CEA $F_E$  34.9%.

Overall, DS remains the weakest performer, which may be due to versioning effects in DeepSeek and its high sensitivity to the specific examples provided in the prompt.

## 5.6 Conclusion

In this study, two coreference resolution models are evaluated (DeepSeek v3 and ChatGPT 4o) under both zero-shot and few-shot prompting across three standard metrics (MUC, B<sup>3</sup>, CEAFF). The main findings can be concluded as:

- **Prompt Structure Matters:** Cluster size and composition strongly influence metric scores. Large, unambiguous clusters boost MUC and B<sup>3</sup>, but CEAFF penalizes even minor errors in large clusters.
- **Dialogue Type Effects:** Pure monologues yield high scores on all metrics. Mixed Q&A and alternating-speaker dialogues introduce missing context and “multiple I” challenges, degrading especially CEAFF performance.
- **Few-Shot Improves Adaptation:** Providing a handful of glossed examples significantly enhances ChatGPT 4o’s performance, particularly in under-resourced German Sign Language scenarios, by clarifying indexing and cluster patterns.
- **Model Comparison:** ChatGPT 4o + few-shot (CF) outperformed all other combinations (MUC 67.7%, B<sup>3</sup> 60.7%, CEAFF 45.3%). DeepSeek v3 + zero-shot (DZ) was second best, suggesting sensitivity to in-context examples. Both zero-shot conditions lag behind, with DeepSeek v3 + few-shot (DS) performing worst overall.
- **Future Directions:** To improve sign-language coreference, future work should explore richer prompt designs for dialogue, more diverse few-shot exemplars, and model adaptations that better handle cluster alignment under CEAFF’s strict matching.

## 5.7 Future Work

Through deep semantic analysis of the original DGS glosses, this study has taken initial steps toward a more fine-grained definition of entity selection logic in DGS. However, as the authors are not domain experts in DGS, the professional validity of some newly introduced entity types—beyond basic pronoun enhancements derived from the original JSON files—remains to be verified in collaboration with native signers and linguists.

This experiments also demonstrated a clear deficiency in current large language models’ ability to resolve coreference in dialogic contexts: none of the models reliably distinguished speaker-specific pronouns when two signers converse simultaneously. Addressing this challenge represents a major avenue for future research.

Looking ahead, i believe that advances in LLM architectures and API-based integrations could substantially improve sign language translation software. At the same time, these case studies highlighted the lack of standardized semantic and syntactic guidelines for gloss annotation. Developing comprehensive annotation schemes and implementing rigorous

manual curation workflows will be critical to enhance LLMs' understanding of coreference phenomena and to elevate the quality of automatic sign language translation.

# Chapter 6

# Appendix

User  
prompt:

You are a sign language coreference resolution assistant.

Input: a list of glosses, each with a unique gloss ID.

1. Identify which glosses refer to entities.
2. Cluster glosses that refer to the same entity.

```
user_prompt = (  
    "$GEST-OFF^*#0 ICH1#1 MG $GEST-OFF^#2  
    ...  
)
```

DS-v3:

==== DS First Reply ===

1. Identify glosses that refer to entities (people, objects, locations, etc.)
2. Group glosses that refer to the same entity into clusters

Here are the identified entity glosses and their clusters:

### Cluster 1: "ICH" (First Person)

- ICH1#1
- ICH2#4
- ICH2#7
- ICH1#12
- ICH1#15
- ICH1#22
- ICH1#29
- ICH1#34
- ICH1#49
- ICH1\*#55

### Cluster 2: "ELTERN" (Parents)

- ELTERN1A#31

### Cluster 3: "TAUB-GEHÖRLOS" (Deaf)

- TAUB-GEHÖRLOS1A\*#10
- TAUB-GEHÖRLOS1A#38
- TAUB-GEHÖRLOS1A\*#46

#...

prompt-1 DS-v3 zero shot

..## Cluster 4: "SCHULE" (School)

- SCHULE3#18
- SCHULE3#25

### Cluster 5: "GEBÄRDEN" (Sign Language)

- GEBÄRDEN1A\*#11

### Cluster 6: "VEREIN" (Association/Club)

- VEREIN2A#39
- VEREIN2A#59

### Cluster 7: "PERSON" (Person)

- PERSON1\*#47

### Cluster 8: "ALL" (Everyone)

- ALLE1A^#52

### Cluster 9: "JUNG" (Young)

- JUNG2#54

### Notes:

1. "AUFWACHSEN1A" (grow up) appears multiple times but is an action, not an entity.
2. "SCHWIMMEN1" (swim) is also an action, not an entity.
3. "ERSTES-MAL3A" (first time) refers to an event/instance, not a concrete entity.
4. "GEHÖRLOSENVEREIN" (deaf club/association) would be part of Cluster 6 if explicitly glossed, but here it seems to be referenced by "TAUB-GEHÖRLOS1A" + "VEREIN2A" in sequence.

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped  
into that cluster."

DS-v3:

==== DS Follow-up Reply ===

Here's the expanded cluster analysis with explanations:

### Cluster 1: First-Person Speaker (ICH)

Gloss IDs: #1, #4, #7, #12, #15, #22, #29, #34, #49, #55

\*\*Description\*\*: All these glosses are variations of "ICH" (German for "I"), representing the signer/narrator throughout the story. They are clustered together because:

- Different inflection markers (ICH1, ICH2) still refer to the same person.
- Even when spatially modified (e.g., ICH1\*#12 with a star marker), the core reference remains the signer.
- Temporal references (e.g., younger vs. current self in #55 vs. #49) still point to the same individual over time.

---

### Cluster 2: Parents (ELTERN)

Gloss ID: #31

\*\*Description\*\*: Explicitly glossed as "ELTERN1A" (parents). This is a singular reference with no other glosses linked to it in this narrative, representing the signer's parental figures.

---

### Cluster 3: Deaf Identity/Community (TAUB-GEHÖRLOS)

Gloss IDs: #10, #38, #46

\*\*Description\*\*: These glosses all use the "TAUB-GEHÖRLOS" (deaf/hard-of-hearing) sign, representing:

- #10: A general identity marker (likely modifying the narrator's upbringing).
  - #38: Specific reference to a "gehörlosenverein" (deaf association).
  - #46: Modified with a star marker (\*), possibly referring to other deaf individuals in the community.
- \*\*Why clustered\*\*: All describe deafness as an identity/group attribute, whether for the narrator or others.

---

prompt-1 DS-v3 zero shot

## DS-v3:

### Cluster 4: School (SCHULE)

Gloss IDs: #18, #25

**\*\*Description\*\*:** Both instances refer to the same institutional context ("school") where the narrator was educated until age 15. The repeated gloss (with different ID #s) refers to the same physical/conceptual place.

---

### Cluster 5: Sign Language (GEBÄRDEN)

Gloss ID: #11

**\*\*Description\*\*:** Standalone reference to "GEBÄRDEN1A" (sign language). This is a unique entity in the narrative, representing the language itself rather than an actor.

---

### Cluster 6: Deaf Association (VEREIN)

Gloss IDs: #39, #59

**\*\*Description\*\*:** Both glosses use "VEREIN2A" (club/association), specifically tied to the deaf community:

- #39: First mention when joining the swimming group.
- #59: Later reference ("not allowed in the association").

**\*\*Why clustered\*\*:** The same institutional entity is referenced across different time points.

---

### Cluster 7: Generic Person (PERSON)

Gloss ID: #47

**\*\*Description\*\*:** "PERSON1\*" refers to an unspecified individual (possibly modified spatially with \*). Isolated reference but treated as a distinct entity.

---

### Cluster 8: Collective "All" (ALLE)

Gloss ID: #52

**\*\*Description\*\*:** "ALLE1A^" (everyone) is a plural entity referring collectively to people in a group (likely the deaf association members). The ^ modifier suggests spatial grouping.

---

prompt-1 DS-v3 zero shot

### Cluster 9: Youth/Jung (JUNG)

Gloss ID: #54

\*\*Description\*\*: "JUNG2" describes a state of being young, but here it contextualizes the narrator as a younger self (linked to time via "VORHER1#53"). Treated as an attribute entity.

--

### Key Observations:

- \*\*Actions vs. Entities\*\*: Actions like "AUFWACHSEN1A" (grow up) or "SCHWIMMEN1" (swim) were excluded as they describe events.
- \*\*Spatial/Modifier Distinction\*\*: Starred/hatted glosses (e.g., \* or ^) were still grouped if their core reference matched (e.g., ICH1\* with ICH1).
- \*\*Temporal Scope\*\*: References like "young me" (#55) were included in the "ICH" cluster despite time differences because they corefer to the narrator.

--

--

User  
prompt:

You are a sign language..(same as zero shot)(input p1)

----- Example -----

Context: A person is talking to someone else about Diana.

Input:

```
$INDEX1#0
$ALPHA1:D-I-A#1
STILL2*#2
TO-SEE1*#3
$INDEX1#4
MEMORY-OR-REMINDER2#5
PAST-OR-BACK-THEN2#6
$GEST-OFF^#7
```

Output (pure JSON):

```
{
  "$INDEX1#0": 0,
  "$ALPHA1:D-I-A#1": 0,
  "$INDEX1#4": 1,
  "MEMORY-OR-REMINDER2#5": 1
}
```

Entity descriptions:

Entity 0: “Diana”

Entity 1: “memory / reminder of the past”

””

DS-v3:

(SEE NEXT)

prompt-1 DS-v3 few shot

DS-v3:

==== DS First Reply ===

```
```json
{
    "$INDEX1#35": 0,
    "$INDEX1#60": 0,
    "$INDEX1#62": 0,
    "ICH1#1": 1,
    "ICH2#4": 1,
    "ICH2#7": 1,
    "ICH1#12": 1,
    "ICH1#15": 1,
    "ICH1#22": 1,
    "ICH1#29": 1,
    "ICH1#34": 1,
    "ICH1#49": 1,
    "ICH1*#55": 1,
    "ELTERN1A#31": 2,
    "$NUM-TEEN1:5#16": 3,
    "$NUM-TEEN1:5#27": 3,
    "SCHULE3#18": 4,
    "SCHULE3#25": 4,
    "TAUB-GEHÖRLOS1A#38": 5,
    "TAUB-GEHÖRLOS1A*#46": 5,
    "TAUB-GEHÖRLOS1A*#10": 5,
    "VEREIN2A#39": 6,
    "VEREIN2A#59": 6
}
```

```

"Please, in English, expand the description for each entity cluster (all entity IDs)."  
"For every entity, explain what it represents and why the glosses were grouped  
into that cluster."

### DS-v3:

==== DS Follow-up Reply ====

Entity descriptions:

- Entity 0: "Index points" - All instances of \$INDEX1 (glosses #35, #60, #62) are clustered here. These likely represent pointing signs used for spatial reference or to indicate specific referents in the discourse.
- Entity 1: "The narrator" - All instances of ICH1/ICH2 (glosses #1, #4, #7, #12, #15, #22, #29, #34, #49, #55) are clustered as they all refer to the signer/narrator of the story (1st person references). The different numeral markings (ICH1 vs ICH2) represent the same referent with different grammatical inflections.
- Entity 2: "Parents" - ELTERN1A#31 is clustered separately as it refers to a distinct entity (the narrator's parents) mentioned in the narrative.
- Entity 3: "The number fifteen" - Both instances of \$NUM-TEEN1:5 (#16, #27) refer to the same numerical value (15) mentioned at different points in the narrative.
- Entity 4: "School" - Both SCHULE3 instances (#18, #25) refer to the same school concept mentioned at different points in the narrative.
- Entity 5: "Deaf/Deafness" - All TAUB-GEHÖRLOS1A instances (#10, #38, #46) refer to the concept of deafness or the deaf community, appearing in different contexts but representing the same core concept.
- Entity 6: "Club/Association" - Both VEREIN2A instances (#39, #59) refer to what appears to be a deaf club/association that is mentioned multiple times in the narrative.

User  
prompt:

You are a sign language coreference resolution assistant.  
Input: a list of glosses, each with a unique gloss ID.  
1. Identify which glosses refer to entities.  
2. Cluster glosses that refer to the same entity  
...(P2\_input)

DS-v3:

== DS First Reply ==

1. **Entity Glosses Identification**:

- Pronouns and indexing: ICH1#55, \$INDEX1#60, \$INDEX1#62, ICH1#79, \$INDEX1#68, \$INDEX1#71, \$INDEX1#84, ICH2#85, ICH2#95, WIR1A#129
- Entities: JUNG2#54 (young person), VEREIN2A#59 (club), ERWACHSENE2A#78 (adult), JUGENDLEITER (youth leader via \$INDEX1#68), JUNG2#69 (young person), JUNG2#89 (young person), JUNG2#113 (young people), ALLE2A^#87 (everyone), TAUB-GEHÖRLOS1A#110 (deaf people), ERWACHSENE2A#117 (adults), MASSE-PERSON3^#122 (group of people), ALLE1A#114 (all)

2. **Coreference Clusters**:

- **Cluster 1 (Speaker)**: ICH1#55, ICH1#79, ICH2#85, ICH2#95, WIR1A#129. These refer to the signer/speaker (first person).
- **Cluster 2 (Youth)**: JUNG2#54, JUNG2#69, JUNG2#89, JUNG2#113. These refer to young people or youth.
- **Cluster 3 (Adults)**: ERWACHSENE2A#78, ERWACHSENE2A#117. These refer to adults.
- **Cluster 4 (Indexed Entities)**:
  - \$INDEX1#60, \$INDEX1#62: Likely the same referent (context suggests club or activity leader).
  - \$INDEX1#68, \$INDEX1#71: Refers to "jugendleiter" (youth leader).
  - \$INDEX1#84: May refer back to the speaker or another participant.
- **Cluster 5 (Groups)**: ALLE2A^#87, ALLE1A#114, MASSE-PERSON3^#122. These refer to groups or everyone.
- **Cluster 6 (Deaf Community)**: TAUB-GEHÖRLOS1A#110. Refers to deaf people.
- **Cluster 7 (Activities/Objects)**: VEREIN2A#59 (club), UNTERHALTUNG2A#64/76/81/86/98/125 (conversation/entertainment), SPASS1#80/103/126 (fun), SCHWIMMEN1#66/119 (swimming). These are non-referential or event-based.

---

prompt-2 DS-v3 zero shot

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped  
into that cluster."

DS-v3:

==== DS Follow-up Reply ===

#### #### \*\*Coreference Clusters\*\*

##### ##### \*\*Cluster 1: Speaker (First Person)\*\*

- \*\*Glosses\*\*: ICH1#55, ICH1#79, ICH2#85, ICH2#95, WIR1A#129
- \*\*Description\*\*: These refer to the \*\*signer (first person)\*\*.
  - \*\*ICH1#55 & ICH1#79\*\*: Both use "ICH" (I), referring to the speaker.
  - \*\*ICH2#85 & ICH2#95\*\*: Variants of first-person reference (possibly different grammatical roles).
  - \*\*WIR1A#129\*\*: "WIR" (we) is an inclusive form, likely extending to the speaker and others (e.g., a group activity).

##### ##### \*\*Cluster 2: Youth / Young People\*\*

- \*\*Glosses\*\*: JUNG2#54, JUNG2#69, JUNG2#89, JUNG2#113
- \*\*Description\*\*: These refer to \*\*young people\*\* (e.g., children, teenagers).
  - \*\*JUNG2#54\*\*: Introduced early ("jung" – young).
  - \*\*JUNG2#69, JUNG2#89, JUNG2#113\*\*: Reiterated later, often in contexts about leadership (e.g., "JUGENDLEITER #68") or group activities (e.g., "JUNG2#113" for youth in a trip).

##### ##### \*\*Cluster 3: Adults\*\*

- \*\*Glosses\*\*: ERWACHSENE2A#78, ERWACHSENE2A#117
- \*\*Description\*\*: Refers to \*\*adults\*\* (contrasting with youth).
  - \*\*ERWACHSENE2A#78\*\*: Introduced with "INTERESSE" (interest), possibly about adult supervision.
  - \*\*ERWACHSENE2A#117\*\*: Later reference, e.g., traveling with adults ("MIT ERWACHSENE").

##### ##### \*\*Cluster 4: Indexed Referents (Points/Contextual Entities)\*\*

- \*\*Glosses\*\*: \$INDEX1#60, \$INDEX1#62, \$INDEX1#68, \$INDEX1#71, \$INDEX1#84
- \*\*Description\*\*: These are \*\*indexing signs\*\* (pointing to entities in signing space).
  - \*\*\$INDEX1#60 & #62\*\*: Likely the same referent, possibly a \*\*club (VEREIN2A#59)\*\* or activity location.
  - \*\*\$INDEX1#68 & #71\*\*: Explicitly linked to "jugendleiter" (youth leader).
  - \*\*\$INDEX1#84\*\*: May refer back to the speaker or another participant (context: "MEHR ERFAHRUNG").

prompt-2 DS-v3 zero shot

### ### \*\*Cluster 5: Groups / Collective Entities\*\*

- \*\*Glosses\*\*: ALLE2A^#87, ALLE1A^#114, MASSE-PERSON3^#122
- \*\*Description\*\*: General references to \*\*groups or everyone\*\*.
- \*\*ALLE2A^#87\*\*: "ALLE" (all) in a context of participation.
- \*\*ALLE1A^#114\*\*: Reiterated in a trip context ("ALLE FAHREN").
- \*\*MASSE-PERSON3^#122\*\*: "Group of people" (e.g., crowds in activities).

### #### \*\*Cluster 6: Deaf Community\*\*

- \*\*Glosses\*\*: TAUB-GEHÖRLOS1A#110
- \*\*Description\*\*: Explicitly refers to \*\*deaf/hard-of-hearing people\*\*.
- Used in "DEUTSCH MEISTERSCHAFT AUCH TAUB-GEHÖRLOS" (German championships for deaf people).

### #### \*\*Cluster 7: Non-Referential / Event-Based Entities\*\*

- \*\*Glosses\*\*:
- \*\*VEREIN2A#59\*\*: (club).
- \*\*UNTERHALTUNG2A#64/76/81/86/98/125\*\*: (conversation/entertainment).
- \*\*SPASS1#80/103/126\*\*: (fun).
- \*\*SCHWIMMEN1#66/119\*\*: (swimming).

- \*\*Description\*\*: These are \*\*activities or objects\*\*, not coreferential entities.  
They describe events/actions rather than specific referents.

### ### \*\*Notes on Ambiguity\*\*

- Some indexing signs (\*\*\$INDEX1#60/62/84\*\*) could link to contextually introduced entities (e.g., club, youth leader).
- \*\*WIR1A^#129\*\* might include the speaker + others but is grouped with first-person references for simplicity.

Let me know if you'd like further refinement!

User  
prompt:

You are a sign language..(same as zero shot)(input p1)  
----- Example -----

Context: A person is talking to someone else about Diana.  
[same example]

DS-v3:

==== DS First Reply ===

```
```json
{
  "$INDEX1#60": 0,
  "ICH1*#55": 1,
  "$INDEX1#62": 2,
  "$INDEX1#68": 3,
  "$INDEX1#71": 4,
  "ICH1#79": 5,
  "$INDEX1#84": 6,
  "ICH2#85": 7,
  "ICH2#95": 8,
  "WIR1A*#129": 9
}
```

Entity descriptions:

Entity 0: "contextual referent indexed by \$INDEX1#60"

Entity 1: "speaker (self-reference) as in ICH1\*#55"

Entity 2: "referenced entity by \$INDEX1#62"

Entity 3: "youth leader referenced by \$INDEX1#68"

Entity 4: "another reference to the youth leader or related entity by \$INDEX1#71"

Entity 5: "speaker again as in ICH1#79"

Entity 6: "referenced entity by \$INDEX1#84"

Entity 7: "different speaker or self-reference variation as in ICH2#85"

Entity 8: "another self-reference or different speaker as in ICH2#95"

Entity 9: "group reference including speaker as in WIR1A\*#129"

```

prompt-2 DS-v3 few shot

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

## DS-v3:

### ==== DS Follow-up Reply ===

Here's the refined JSON with detailed explanations for each entity cluster:

```
    `` ` json
{
  "$INDEX1#60": 0,
  "ICH1*#55": 1,
  "$INDEX1#62": 2,
  "$INDEX1#68": 3,
  "$INDEX1#71": 4,
  "ICH1#79": 5,
  "$INDEX1#84": 6,
  "ICH2#85": 7,
  "ICH2#95": 8,
  "WIR1A*#129": 9,
  "ERWACHSENE2A",
  "ERWACHSENE2A",
  "JUNG2#54": 11,
  "JUNG2#69": 11,
  "JUNG2#89": 11,
  "JUNG2#113": 11,
  "TAUB-GEHÖRLOS
}
```

## DS-v3:

Entity descriptions:

Entity 0: "Unknown third-person referent established at position #60 (likely a person being discussed)"

Entity 1: "The primary signer/speaker (first-person reference), marked by the first occurrence of ICH1\* at position #55"

Entity 2: "A distinct third-person referent established at position #62 (possibly a different person from Entity 0)"

Entity 3: "Youth leader mentioned at position #68, distinguished by the specific role label 'jugendleiter'"

Entity 4: "Potential co-reference or related person to the youth leader (position #71), possibly the same as Entity 3 but with insufficient context to confirm"

Entity 5: "The signer/speaker again (first-person reference) at position #79"

Entity 6: "Another distinct third-person referent established at position #84"

Entity 7: "Alternate first-person reference or possibly different speaker using ICH2 variation at position #85"

Entity 8: "Additional first-person reference at position #95, potentially same as Entity 7 or another speaker"

Entity 9: "Group reference ('we') including the signer, occurring at position #129"

Entity 10: "Adults group (ERWACHSENE2A), clustered from positions #78 and #117 referring to the same category"

Entity 11: "Young people/youth group (JUNG2), clustered from multiple positions (#54, #69, #89, #113) with consistent reference"

Entity 12: "Deaf/hard-of-hearing community reference (TAUB-GEHÖRLOS1A) at position #110"

}

Key observations:

1. First-person references (ICH/ICH2) are clustered separately as they may represent different speakers or morphological variations
2. Numbered index signs (\$INDEX) are kept distinct unless clear co-reference is evident
3. Role-labeled entities (like "jugendleiter") get their own clusters
4. Group nouns (ERWACHSENE, JUNG) are clustered by semantic category despite position distance
5. The deaf community reference is separated as it's a distinct social category

Note: Some clusters could potentially be merged with more contextual evidence (e.g., Entities 3-4 or 7-8), but are kept separate here due to insufficient co-reference markers in the input.

## ## Step 1: Identify Entity-Referring Glosses

From the provided list, the following glosses likely refer to entities:

- \*\*[[ICH1]]\*\* (ID: 135) - First-person singular pronoun ("I")
- \*\*[[ICH2]]\*\* ich (ID: 147) - First-person singular pronoun ("I")
- \*\*MITGLIED1\*\* mitglied (ID: 138) - Refers to a "member" (likely of a club or group)
- \*\*KNIE1B\*\* knie (ID: 148) - Body part ("knee")

## ### Step 2: Coreference Clustering

Here are the coreference clusters for the identified entity-referring glosses:

### 1. \*\*Cluster 1 (First-Person Speaker):\*\*

- [[ICH1]] (ID: 135)
- [[ICH2]] ich (ID: 147)

\*(These are clustered together because they both refer to the same speaker in different instances.)\*

### 2. \*\*Cluster 2 (Other Entities):\*\*

- MITGLIED1 mitglied (ID: 138) - \*Refers to a "member" (likely distinct from the speaker)\*
- KNIE1B knie (ID: 148) - \*Refers to a body part, not coreferent with others in this list\*

## ### Notes:

- \*\*KEGELN1/KEGELN1\*\*\* (IDs: 136, 139, 143) refer to the activity "kegeln" (bowling) but do not denote entities—they are action glosses.
- \*\*JAHR1B\*\*\* (ID: 137) refers to "ten years," which is a temporal expression, not an entity.
- The rest are non-entity function words (e.g., "dann," "später," "um," "bis") or gestures (\*\*\$GEST-OFF^\*\*).

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

DS-v3:

==== DS Follow-up Reply ====

### #### \*\*Clustered Entities and Their Descriptions:\*\*

#### #### \*\*1. First-Person Speaker Cluster ([[ICH1]], [[ICH2]])\*\*

- \*\*[[ICH1]] (ID: 135):\*\*

- A singular first-person pronoun ("I"). This typically refers to the signer themselves in a narrative or conversational context.

- The \*\*double brackets\*\* (`[[ ]]`) conventionally indicate a constructed or modified sign, often used for pronouns or named entities in sign language glossing.

- \*\*[[ICH2]] ich (ID: 147):\*\*

- A repetition or variant of the first-person pronoun ("I"), possibly used later in the discourse (e.g., switching back to the speaker's perspective after discussing other topics).

- \*\*Reason for Clustering:\*\*

These two glosses are grouped together because both refer to the same entity—the speaker—in different parts of the discourse. Despite slight variations (e.g., bracket notation, presence/absence of "ich"), their semantic and pragmatic functions are identical.

#### #### \*\*2. Distinct Referents (MITGLIED1, KNIE1B)\*\*

- \*\*MITGLIED1 mitglied (ID: 138):\*\*

- Refers to a "member" (e.g., of a club, team, or organization). This is an entity but does not corefer with any other gloss in the provided input.

- Likely introduced as a new referent (e.g., "I was a member of a bowling club").

- \*\*KNIE1B knie (ID: 148):\*\*

- Refers to the body part "knee." This is an inanimate object and does not corefer with other glosses in this context.

- Might appear in a phrase like "my knee hurt" or "I injured my knee."

- \*\*Reason for Non-Clustering:\*\*

These two glosses are \*not\* grouped together because they represent entirely distinct entities (a person's role vs. a body part) with no semantic or discourse links in the given input.

### ### \*\*Excluded Glosses (Non-Entities):\*\*

The following were \*\*not\*\* treated as entities because they denote actions, temporal expressions, or function words:

- \*\*KEGELN1/KEGELN1\*\*\* (IDs: 136, 139, 143): Repeated references to the activity "bowling."

- \*\*JAHR1B\*\*\* (ID: 137): Temporal phrase "ten years."

- \*\*DANN1B\*, SPÄTER10\*, UM2A, etc.:\*\* Adverbial or connective words (e.g., "then," "later," "around").

- \*\*\$GEST-OFF^\*\*\* (ID: 151): A gesture marker, not a referential entity.

prompt-3 DS-v3 zero shot

User  
prompt:

You are a sign language..(same as zero shot)(input p1)

----- Example -----

Context: A person is talking to someone else about Diana.  
[same example]

DS-v3:

==== DS First Reply ===

```
{  
  "[[ICH1]] #135": 0,  
  "[[ICH2]] ich #147": 0,  
  "JAHR1B* zehn jahre #137": 1,  
  "KEGELN1* #143": 2  
},
```

Entity descriptions:

Entity 0: "The signer (first person)"

Entity 1: "The duration of ten years"

Entity 2: "The activity of bowling"

prompt-3 DS-v3 few shot

==== DS Follow-up Reply ====

about that example as well.

Here's the corrected and expanded version of your response:

```
```json
{
  "$INDEX1#0": 0,
  "$ALPHA1:D-I-A#1": 0,
  "$INDEX1#4": 1,
  "MEMORY-OR-REMINDER2#5": 1
}
```

```

Entity descriptions:

- **Entity 0**: "Diana" - This cluster represents the proper name "Diana" spelled out (\$ALPHA1:D-I-A#1) and then later referenced using an indexical sign (\$INDEX1#0). In sign languages, fingerspelling a name and then later pointing to the same location in signing space is a common way to maintain reference to a person.

- **Entity 1**: "memory/reminder of the past" - This cluster contains both the sign for "memory" (MEMORY-OR-REMINDER2#5) and a subsequent indexical reference (\$INDEX1#4) that points back to that memory concept. The cluster represents the abstract concept of something remembered from the past, where the pointing sign refers back to the previously established memory concept.

Please note that for the second input you provided (the German bowling example), you already included entity descriptions, though it appears you might have been asking

User  
prompt:

You are a sign language coreference resolution assistant.  
Input: a list of glosses, each with a unique gloss ID.  
1. Identify which glosses refer to entities.  
2. Cluster glosses that refer to the same entity  
...(P4\_input)

### DS-v3:

==== DS First Reply ===

1. First, identify the entity-referring glosses (based on conceptual, not formal, criteria).

Entity-referring glosses are those that denote specific entities like people (e.g., pronouns, names), places, events, or objects.

- `ICH1` , `ICH2` (variations of "I" – the signer)
- `HAMBURG1` (place)
- `DEUTSCH1 MEISTERSCHAFT1` (event)
- `GRENZE1A` (conceptual entity – "limit")
- `STARTSCHUSS1` (event)
- `ZIEL4` (conceptual entity – "goal/finish line")
- `ZUSAMMEN-PERSON1` (people – "others/competitors")
- `ERFOLG2` (conceptual entity – "success")
- `PLATZ9` (conceptual entity – "place/ranking")

2. Clustering glosses that refer to the same entity:

- \*\*Signer (ICH1/ICH2)\*\*: #153, #157, #164, #167, #171, #177, #179, #181, #188, #193, #199
- \*\*Hamburg (place)\*\*: #155, #158
- \*\*Event (German Championship)\*\*: #161, #162, #163
- \*\*Limit/GRENZE\*\*: #172
- \*\*Start signal/STARTSCHUSS\*\*: #174
- \*\*Goal/ZIEL\*\*: #184
- \*\*Others/ZUSAMMEN-PERSON\*\*: #186
- \*\*Success/ERFOLG\*\*: #213
- \*\*Place/PLATZ (ranking)\*\*: #212

Note:

- Some glosses (e.g., `DORT1`) are locative but do not denote standalone entities.
- Gestures (`\$GEST-OFF`, `\$INDEX1`) are not entities.
- Conceptual entities like `ATEMNOT1` ("shortness of breath") describe states rather than referential entities.

The signer is the central entity, with clusters reflecting self-reference and event participation.

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

DS-v3:

==== DS Follow-up Reply ====

Here are the expanded descriptions for each entity cluster:

1. \*\*Signer (ICH1/ICH2)\*\*

- Gloss IDs: #153, #157, #164, #167, #171, #177, #179, #181, #188, #193, #199
- Explanation: These all represent the first-person perspective (the narrator/signer). Variations like ICH1/ICH2 are morphological differences (e.g., directionality, repetition) but refer to the same entity. The clustering reflects consistent self-reference throughout the narrative, whether describing actions (e.g., swimming, running) or states (e.g., exhaustion).

2. \*\*Hamburg (place)\*\*

- Gloss IDs: #155, #158
- Explanation: Both glosses denote the city of Hamburg as a location. The repetition (with slight formal variation) indicates spatial anchoring in the narrative—likely the setting for the described swimming competition.

3. \*\*German Championship (event)\*\*

- Gloss IDs: #161 (KAMPF1A), #162 (DEUTSCH1), #163 (MEISTERSCHAFT1)
- Explanation: This cluster forms a coherent event entity. "DEUTSCH1 MEISTERSCHAFT1" (German Championship) is modified by "KAMPF1A" (struggle/competition), collectively referring to the swimming championship context. The grouping captures the event's institutional and competitive aspects.

4. \*\*Limit (GRENZE1A)\*\*

- Gloss ID: #172
- Explanation: This stands for a conceptual boundary—likely the narrator's physical or mental limit during the race. It's singled out as an entity because it's a pivotal narrative element (e.g., "pushing past limits").

5. \*\*Start signal (STARTSCHUSS1)\*\*

- Gloss ID: #174
- Explanation: Represents the auditory/visual signal initiating the race. As a discrete event marker, it's crucial for sequencing the narrative's action.

6. \*\*Goal/finish line (ZIEL4)\*\*

- Gloss ID: #184
- Explanation: The endpoint of the race. Though mentioned once, it's a key spatial and temporal entity structuring the competition narrative.

prompt-4 DS-v3 zero shot

7. **Other competitors (ZUSAMMEN-PERSON1)**

- Gloss ID: #186
- Explanation: Refers to fellow swimmers as a collective entity. The gloss implies a group dynamic (e.g., close competition), justifying its separation from individual actors.

8. **Success (ERFOLG2)**

- Gloss ID: #213
- Explanation: An abstract but narratively salient entity—the outcome (placing second) framed as achievement.

9. **Place/ranking (PLATZ9)**

- Gloss ID: #212
- Explanation: Specifically denotes the narrator's final rank ("zweite PLATZ9" – second place). It's grouped separately from generic "success" as it quantifies the outcome.

---

**Exclusion Notes:**

- Gestures (e.g., `\$GEST-OFF`) and locatives (e.g., `DORT1`) were excluded as they lack referentiality.
- States like `ATEMNOT1` (shortness of breath) describe conditions rather than entities.
- Repeated signs (e.g., `SCHWIMMEN1`) denote actions, not coreferential entities.

This clustering prioritizes narratively significant entities with clear referential continuity.

User  
prompt:

You are a sign language..(same as zero shot)(input p1)

----- Example -----

Context: A person is talking to someone else about Diana.  
[same example]

DS-v3:

==== DS First Reply ===

```
```json
{
  "ICH1#157": 0,
  "ICH1#164": 0,
  "ICH1#171": 0,
  "ICH1#181": 0,
  "ICH1#188": 0,
  "ICH1#193": 0,
  "ICH1#199": 0,
  "ICH2#153": 0,
  "ICH2#167": 0,
  "ICH2*#177": 0,
  "ICH2*#179": 0,
  "HAMBURG1#155": 1,
  "HAMBURG1*#158": 1,
  "DORT1#156": 2,
  "DORT1#159": 2,
  "DORTHIN-GEHEN1*#165": 2,
  "SCHWIMMEN1*#160": 3,
  "SCHWIMMEN1*#166": 3,
  "SCHWIMMEN1#169": 3,
  "SCHWIMMEN1 schwimmen#219": 3,
  "SCHWIMMEN1 schwimmen#222": 3,
  "SCHWIMMEN1 schwimmen#228": 3,
  "DEUTSCH1#162": 4,
  "MEISTERSCHAFT1#163": 5,
  "ZIEL4#184": 6,
  "PLATZ9 platz#212": 7,
  "ERFOLG2 erfolg#213": 8
}
```

prompt-4 DS-v3 few shot

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs)."  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

DS-v3:

==== DS Follow-up Reply ====

```
```json
{
  "entity_descriptions": {
    "0": "The speaker, referring to themselves multiple times with variations of 'ICH1' and 'ICH2' (e.g., 'ICH1#157', 'ICH1#164', 'ICH2#153', etc.). These are grouped together as they all represent the first-person perspective in the narrative, indicating the speaker's actions or state.",  

    "1": "The city of Hamburg, mentioned in 'HAMBURG1#155' and 'HAMBURG1*#158'. These glosses are clustered because they directly refer to the same location where the events took place.",  

    "2": "References to a specific location or direction, such as 'DORT1#156', 'DORT1#159', and 'DORTHIN-GEHEN1*#165'. These are grouped because they all indicate a place or movement toward a place within the context of the narrative (likely related to Hamburg or the swimming competition).",  

    "3": "The activity of swimming, represented by 'SCHWIMMEN1*#160', 'SCHWIMMEN1*#166', 'SCHWIMMEN1#169', etc. These are grouped as they all refer to the same action, which is central to the story (a swimming competition).",  

    "4": "The German national context, represented by 'DEUTSCH1#162'. This stands alone as it refers to the nationality or language associated with the competition.",  

    "5": "The championship or master-level competition, represented by 'MEISTERSCHAFT1#163'. This is unique as it refers to the tier or level of the swimming event.",  

    "6": "The goal or finish line, represented by 'ZIEL4#184'. This refers to the endpoint of the race or competition.",  

    "7": "The placement or ranking, represented by 'PLATZ9 platz#212'. This refers to the speaker's position in the competition (e.g., second place).",  

    "8": "The achievement or success, represented by 'ERFOLG2 erfolg#213'. This stands alone as it refers to the outcome or result of the competition."
  },
  "additional_notes": "Some glosses (e.g., 'GRENZE1A^*#172', 'STARTSCHUSS1*#174', 'ATEMNOST1#194') were not included in the clusters because they describe actions, events, or states (e.g., 'falling', 'running', 'shortness of breath') rather than referring to distinct entities. The focus was on clustering references to people, places, and concrete concepts."
}
```

prompt-4 DS-v3 few shot

User  
prompt:

You are a sign language coreference resolution assistant.

Input: a list of glosses, each with a unique gloss ID.

1. Identify which glosses refer to entities.
  2. Cluster glosses that refer to the same entity
- ...(P5\_input)

DS-v3:

==== DS First Reply ====

1. First talk about how to identify which glosses refer to entities.

First, I will identify glosses that refer to entities. In sign language, entities can be referred to by:

- Pronouns (e.g., ICH1, DU1\*)
- Index signs (\$INDEX1) that point to locations in signing space
- Nouns that refer to concrete things (e.g., WASSER1, LAND1A\*)
- Proper nouns (none in this example)
- Gestures that represent entities (\$GEST-...)

Some glosses are not entities:

- Verbs (e.g., SCHWIMMEN1)
- Adjectives (e.g., FRISCH1)
- Adverbs (e.g., IMMER4A\*)
- Discourse markers (e.g., \$ORAL^)
- Numbers (e.g., \$NUM-ZEHN1:2d\*)

2. Now, cluster the glosses that refer to the same entity.

Based on the conversation, here are the coreference clusters:

1) Speaker B:

- DU1\*#546, DU1\*#553, DU1\*#562

2) Speaker A:

- ICH1#238, ICH1\*#242, ICH1\*#245, ICH1\*#257
- ICH2#259, ICH2#587, ICH2#589

3) Water:

- WASSER1#247, WASSER11#581

4) Land (as opposed to water):

- LAND1A\*#574

5) Sport:

- SPORT4B#251

prompt-5 DS-v3 zero shot

## DS-v3:

6) Competition:

- KAMPF1A#240, KAMPF1A#272

7) Fun/Enjoyment:

- SPASS1#246, SPASS1\*#571

8) Private/Free time:

- PRIVAT1A#265, FREI1#266, FREIZEIT1B\*#267

Note that index signs (\$INDEX1) would typically refer to other entities already mentioned, but without visual information about their pointing locations, we can't definitively assign them to clusters. Also, some signs like SCHLUSS1 appear both as a concept and as part of an utterance (e.g., SCHLUSS1#552, SCHLUSS1#273), but they don't refer to specific entities.

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

DS-v3:

==== DS Follow-up Reply ====

Also note possible relations to other entities, if any.

**\*\*Coreference Clusters Analysis:\*\***

1. **\*\*Speaker B (DU1\* cluster)\*\***

- Gloss IDs: #546, #553, #562
- Explanation: All these refer to the second person singular "you" directed at Speaker B throughout the conversation. The asterisk (\*) in the gloss notation typically indicates indexing in signing space, confirming these all refer to the same entity. This cluster represents Speaker A addressing their conversational partner.

2. **\*\*Speaker A (ICH1/ICH2 cluster)\*\***

- Gloss IDs: #238, #242, #245, #257, #259, #587, #589
- Explanation: These represent first person references by Speaker A. While ICH1 and ICH2 are different signs (likely different forms of "I"), they corefer to the same entity - the signer themselves. The cluster shows Speaker A referring to themselves at different points, discussing their swimming preferences, energy levels, and training routines.

3. **\*\*Water References\*\***

- Gloss IDs: #247, #581
- Explanation: These both clearly refer to the concept of water (WASSER1 and WASSER11 with slight variation). They appear in contexts discussing swimming and preferences between water and land. Related to LAND1A\* (#574) which contrasts with water in the discussion.

4. **\*\*Land Reference\*\***

- Gloss ID: #574
- Explanation: This singular reference to "land" appears in contrast to water, discussing where Speaker A prefers to train. The asterisk suggests spatial indexing. This is conceptually related to but distinct from the water cluster.

5. **\*\*Sport Concept\*\***

- Gloss ID: #251
- Explanation: This standalone reference to "sport" represents the general concept. It relates to clusters about swimming and competition but is more generic in nature.

**6. \*\*Competition References\*\***

- Gloss IDs: #240, #272
- Explanation: These refer to "wettkampf" (competition) in German Sign Language.

While appearing in different contexts (one general reference, one about concluding discussion of competition), they corefer to the concept of athletic competition. Related to the SPORT cluster but more specific.

**7. \*\*Fun/Enjoyment References\*\***

- Gloss IDs: #246, #571
- Explanation: These refer to the concept of fun/spaß. While one is from Speaker A and one from Speaker B (with dereferencing), they both refer to the same abstract concept. Related to discussion of swimming preferences.

**8. \*\*Private/Free Time Concepts\*\***

- Gloss IDs: #265, #266, #267
- Explanation: This cluster groups concepts about private time and leisure (PRIVAT1A, FREI1, FREIZEIT1B\*). While not strictly coreferential (they reference slightly different aspects), they all point to the same general concept of non-competitive time. Related to the contrast with competition references.

**\*\*Notable Observations:\*\***

- The \$INDEX1 signs (#248, #254, #582) are notable omissions from clusters - these would reference previously established entities based on their spatial indexing which isn't visible in transcript form.
- Some signs like ALT8B (#551) and ALT5A (#237) might seem like candidates for a "age" cluster, but in context they appear adjectival rather than referring to a specific age entity.
- The conversation shows several contrasts between entities (water vs. land, competition vs. free time) that create meaningful relationships between clusters.

## DS-v3:

==== DS First Reply ====

Output (pure JSON):

```
```json
{
  "$INDEX1#248": 0,
  "$INDEX1#582": 0,
  "$INDEX1#585": 0,
  "DU1*#546": 1,
  "DU1*#553": 1,
  "DU1*#562": 1,
  "ICH1#238": 2,
  "ICH1#559": 2,
  "ICH1#572": 2,
  "ICH1#573": 2,
  "ICH1#577": 2,
  "ICH1*#242": 2,
  "ICH1*#245": 2,
  "ICH1*#257": 2,
  "ICH2#259": 3,
  "ICH2#587": 3,
  "ICH2#589": 3,
  "WASSER1#247": 4,
  "WASSER11#581": 4,
  "KAMPF1A#567": 5,
  "KAMPF1A wettkampf#240": 5,
  "KAMPF1A wettkampf #272": 5,
  "SCHWIMMEN1#554": 6,
  "SCHWIMMEN1#558": 6,
  "SCHWIMMEN1*#563": 6,
  "SCHWIMMEN1 schwimmen#255": 6,
  "SCHWIMMEN1 schwimmen gerne #268": 6,
  "HIN-UND-HER1#557": 7,
  "HIN-UND-HER1#570": 7,
  "HIN-UND-HER1*#564": 7,
  "HIN-UND-HER1*#243": 7
}
```

Entity descriptions:

- Entity 0: "Index referring to an entity (e.g., a place or thing)"
- Entity 1: "The listener or addressee (you)"
- Entity 2: "The speaker (I)"
- Entity 3: "The speaker in a different context or role (I2)"
- Entity 4: "Water"
- Entity 5: "Competition or competition context"
- Entity 6: "Swimming"
- Entity 7: "Back and forth movement or action"

You are a sign language..(same as zero shot)(input p1)  
----- Example -----

Context: A person is talking to someone else about  
Diana.  
[same example]

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs)."  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

DS-v3:

```
== DS Follow-up Reply ==
```json
{
  "entity_descriptions": [
    {
      "id": 0,
      "description": "This cluster represents an indexed entity, likely a place or thing, referred to by the signer. The glosses $INDEX1#248, $INDEX1#582, and $INDEX1#585 all use the index sign to point to or reference this entity, which is why they are grouped together. The exact identity of this entity isn't specified in the text, but it's something being referenced multiple times."
    },
    {
      "id": 1,
      "description": "This cluster represents the listener or addressee (you) in the conversation. The glosses DU1*#546, DU1*#553, and DU1*#562 all directly refer to the second person ('you'), which is why they are grouped together. These are instances where the signer is addressing or referring to the other person in the conversation."
    },
    {
      "id": 2,
      "description": "This cluster represents the speaker (I) in the conversation. The glosses ICH1#238, ICH1#559, ICH1#572, ICH1#573, ICH1#577, ICH1*#242, ICH1*#245, and ICH1*#257 all represent first-person references, where the signer is talking about themselves. The variations (with and without *) likely represent different instances or emphases of self-reference."
    }
}
```

prompt-5 DS-v3 few shot

## DS-v3:

```
{  
  "id": 3,  
  "description": "This cluster represents the speaker in a different context or role (I2). The glosses ICH2#259, ICH2#587, and ICH2#589 use a variant form (ICH2) which might indicate the speaker in a different role or context compared to the standard ICH1 form. This could represent a different aspect of the speaker's identity or perspective."  
},  
{  
  "id": 4,  
  "description": "This cluster represents the concept of water. The glosses WASSER1#247 and WASSER11#581 clearly refer to water, with WASSER11 possibly being a variant or modified form. Both signs relate to the same physical substance, which is why they're grouped together."  
},  
{  
  "id": 5,  
  "description": "This cluster represents the concept of competition or competition context. The glosses KAMPF1A#567, KAMPF1A wettkampf#240, and KAMPF1A wettkampf #272 all relate to competitive situations or contexts (note 'wettkampf' is German for 'competition'). The variations in notation likely represent different instances or phrasings of the same core concept."  
},  
{  
  "id": 6,  
  "description": "This cluster represents the action of swimming. The glosses SCHWIMMEN1#554, SCHWIMMEN1#558, SCHWIMMEN1*#563, SCHWIMMEN1 schwimmen#255, and SCHWIMMEN1 schwimmen gerne #268 all clearly relate to swimming, with some variations indicating enjoyment or specific instances of the action. The core concept remains the same, hence the grouping."  
},  
{  
  "id": 7,  
  "description": "This cluster represents back and forth movement or action. The glosses HIN-UND-HER1#557, HIN-UND-HER1#570, HIN-UND-HER1*#564, and HIN-UND-HER1*#243 all represent the concept of going back and forth (note 'HIN-UND-HER' is German for 'back and forth'). The variations indicate different instances or emphases of this movement concept."
```

User  
prompt:

You are a sign language coreference resolution assistant.  
Input: a list of glosses, each with a unique gloss ID.  
1. Identify which glosses refer to entities.  
2. Cluster glosses that refer to the same entity  
...(P1\_input)

GPT-4o:

- Assistant Answer: 1. Identify glosses that refer to entities:
  - - ICH1#1, ICH2#4, ICH2#7, ICH1\*#12, ICH1#15, ICH1#22, ICH1#29, ICH1#34, ICH1#49,
  - ICH1\*#55: These glosses refer to the first person pronoun "I" or "me."
  - - ELTERN1A#31: Refers to "parents."
  - - TAUB-GEHÖRLOS1A\*#10, TAUB-GEHÖRLOS1A#38, TAUB-GEHÖRLOS1A\*#46: Refers to "deaf"
  - or "deaf people."
  - - PERSON1\*#47: Refers to "person" or "people."
  - - VEREIN2A#39, VEREIN2A#59: Refers to "club" or "association."
  -
- 2. Cluster glosses that refer to the same entity:
  - - Cluster 1 (First Person Pronoun): ICH1#1, ICH2#4, ICH2#7, ICH1\*#12, ICH1#15, ICH1#22, ICH1#29, ICH1#34, ICH1#49, ICH1\*#55
  - - Cluster 2 (Parents): ELTERN1A#31
  - - Cluster 3 (Deaf/Deaf People): TAUB-GEHÖRLOS1A\*#10, TAUB-GEHÖRLOS1A#38, TAUB-GEHÖRLOS1A\*#46
  - - Cluster 4 (Person/People): PERSON1\*#47
  - - Cluster 5 (Club/Association): VEREIN2A#39, VEREIN2A#59

prompt-1 GPT4o zero shot

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "

"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

Here is an expanded description for each entity cluster:

1. **Cluster 1 (First Person Pronoun):**

- **Gloss IDs:** ICH1#1, ICH2#4, ICH2#7, ICH1\*#12, ICH1#15, ICH1#22, ICH1#29, ICH1#34, ICH1#49, ICH1\*#55

- **Description:** This cluster represents the first person pronoun "I" or "me."

These glosses are grouped together because they all refer to the speaker or signer themselves. In sign language glossing, "ICH" is used to denote the first person singular, and the various instances (ICH1, ICH2) indicate the same entity being referred to throughout the discourse.

2. **Cluster 2 (Parents):**

- **Gloss IDs:** ELTERN1A#31

- **Description:** This gloss represents "parents." It is a singular entity in this context, referring to the speaker's parents. The gloss "ELTERN" is used to denote this familial relationship.

3. **Cluster 3 (Deaf/Deaf People):**

- **Gloss IDs:** TAUB-GEHÖRLOS1A\*#10, TAUB-GEHÖRLOS1A#38, TAUB-GEHÖRLOS1A\*#46

- **Description:** This cluster represents "deaf" or "deaf people." The gloss "TAUB-GEHÖRLOS" is used to refer to individuals who are deaf or hard of hearing. These glosses are grouped together because they consistently refer to the same community or characteristic throughout the discourse.

## GPT-4o:

4. \*\*Cluster 4 (Person/People):\*\*

- \*\*Gloss IDs:\*\* PERSON1#47

- \*\*Description:\*\* This gloss represents "person" or "people." It is used to refer to individuals in a general sense. In this context, it may refer to people within the deaf community or those involved in the narrative.

5. \*\*Cluster 5 (Club/Association):\*\*

- \*\*Gloss IDs:\*\* VEREIN2A#39, VEREIN2A#59

- \*\*Description:\*\* This cluster represents a "club" or "association." The gloss "VEREIN" is used to denote an organized group or club, likely related to the deaf community or activities mentioned in the narrative. These glosses are grouped together

because they refer to the same entity or organization within the context.

User  
prompt:

You are a sign language coreference resolution assistant...  
(with same example.)

---

GPT-4o:

Assistant 回答: ```json

```
{  
    "ICH1#1": 0,  
    "ICH2#4": 0,  
    "ICH2#7": 0,  
    "ICH1*#12": 0,  
    "ICH1#15": 0,  
    "ICH1#22": 0,  
    "ICH1#29": 0,  
    "ICH1#34": 0,  
    "ICH1#49": 0,  
    "ICH1*#55": 0,  
    "ELTERN1A#31": 1,  
    "$INDEX1#35": 1,  
    "$INDEX1#60": 1,  
    "$INDEX1#62": 1,  
    "SCHULE3#18": 2,  
    "SCHULE3#25": 2,  
    "VEREIN2A#59": 3,  
    "VEREIN2A#39": 3  
}  
```
```

Entity descriptions:

- Entity 0: "the speaker"
- Entity 1: "the speaker's parents"
- Entity 2: "school"
- Entity 3: "deaf club"

prompt-1 GPT4o few shot

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs)."  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

round 2: Entity 0: "the speaker"

- This entity represents the person who is narrating the story. The glosses "ICH1", "ICH2", and "ICH1\*" are all first-person pronouns in German Sign Language, indicating that the speaker is referring to themselves. These glosses are grouped together because they all denote the same individual who is recounting their experiences.

Entity 1: "the speaker's parents"

- This entity refers to the parents of the speaker. The gloss "ELTERN1A" directly translates to "parents" in German Sign Language. The glosses "\$INDEX1#35", "\$INDEX1#60", and "\$INDEX1#62" are indexical signs that point to previously mentioned entities, in this case, the parents. These are grouped together because they all refer to the speaker's parents.

Entity 2: "school"

- This entity represents the concept of "school" in the narrative. The glosses "SCHULE3" are used to denote the school that the speaker attended. These are grouped together because they consistently refer to the same educational institution in the context of the speaker's life story.

Entity 3: "deaf club"

- This entity refers to a club or association for deaf individuals. The gloss "VEREIN2A" translates to "club" or "association," and in this context, it is specified as a "deaf club" with the gloss "TAUB-GEHÖRLOS1A" indicating "deaf." These glosses are grouped together because they collectively describe the same organization that the speaker is discussing.

User  
prompt:

You are a sign language coreference resolution assistant.  
Input: a list of glosses, each with a unique gloss ID.  
1. Identify which glosses refer to entities.  
2. Cluster glosses that refer to the same entity  
...(P2\_input)

GPT-4o:

Assistant : \*\*Entities and Clusters:\*\*

1. \*\*ICH (I/me)\*\*

- ICH1#55
- ICH1#79
- ICH2#85
- ICH2#95

2. \*\*JUNG (young/youth)\*\*

- JUNG2#54
- JUNG2#69
- JUNG2#89
- JUNG2#113 (youth/young people)

3. \*\*VEREIN (club/association)\*\*

- VEREIN2A#59

4. \*\*JUGENDLEITER (youth leader)\*\*

- \$INDEX1#68 (referring to "jugendleiter")

5. \*\*ERWACHSENE (adult)\*\*

- ERWACHSENE2A#78
- ERWACHSENE2A#117

6. \*\*TAUB-GEHÖRLOS (deaf)\*\*

- TAUB-GEHÖRLOS1A#110

prompt-2 GPT4o zero shot

## GPT-4o:

7. \*\*WIR (we)\*\*

- WIR1A#129

8. \*\*ZUG (train)\*\*

- ZUG2A#124

9. \*\*MASSE-PERSON (group of people)\*\*

- MASSE-PERSON3^#122

Note: Some glosses like \$INDEX1 and \$GEST-OFF^\* are placeholders or gestures that may refer to entities in context but are not explicitly named entities themselves.

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs)."  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

Here's an expanded description for each entity cluster:

1. \*\*ICH (I/me)\*\*

- \*\*Gloss IDs:\*\* ICH1#55, ICH1#79, ICH2#85, ICH2#95

- \*\*Description:\*\* This cluster represents the first-person singular pronoun "I" or "me," referring to the speaker or signer themselves. The glosses are grouped together because they all denote the same entity, the person who is communicating the message.

2. \*\*JUNG (young/youth)\*\*

- \*\*Gloss IDs:\*\* JUNG2#54, JUNG2#69, JUNG2#89, JUNG2#113

- \*\*Description:\*\* This cluster refers to the concept of "young" or "youth." The glosses are grouped together because they all describe young people or the state of being young. JUNG2#113 specifically refers to "youth" or "young people" as a group.

3. \*\*VEREIN (club/association)\*\*

- \*\*Gloss IDs:\*\* VEREIN2A#59

- \*\*Description:\*\* This gloss represents a "club" or "association." It is a standalone entity in this context, referring to an organized group or society.

4. \*\*JUGENDLEITER (youth leader)\*\*

- \*\*Gloss IDs:\*\* \$INDEX1#68

- \*\*Description:\*\* This gloss refers to a "youth leader." The use of \$INDEX1 indicates a specific reference to a person who holds the position of leading or guiding young people.

prompt-2 GPT4o zero shot

## GPT-4o:

### 5. \*\*ERWACHSENE (adult)\*\*

- \*\*Gloss IDs:\*\* ERWACHSENE2A\*#78, ERWACHSENE2A\*#117
- \*\*Description:\*\* This cluster represents "adults." The glosses are grouped together because they both refer to mature individuals, distinct from the youth or young people mentioned in other glosses.

### 6. \*\*TAUB-GEHÖRLOS (deaf)\*\*

- \*\*Gloss IDs:\*\* TAUB-GEHÖRLOS1A#110
- \*\*Description:\*\* This gloss refers to "deaf" individuals. It is a standalone entity in this context, representing people who are part of the deaf community.

### 7. \*\*WIR (we)\*\*

- \*\*Gloss IDs:\*\* WIR1A#129
- \*\*Description:\*\* This gloss represents the first-person plural pronoun "we," indicating a group that includes the speaker or signer. It is a standalone entity in this context, referring to a collective group.

### 8. \*\*ZUG (train)\*\*

- \*\*Gloss IDs:\*\* ZUG2A#124
- \*\*Description:\*\* This gloss represents a "train." It is a standalone entity in this context, referring to a mode of transportation.

### 9. \*\*MASSE-PERSON (group of people)\*\*

- \*\*Gloss IDs:\*\* MASSE-PERSON3^#122
- \*\*Description:\*\* This gloss refers to a "group of people." It is a standalone entity in this context, representing a collective of individuals, possibly in a social or organized setting.

Each cluster groups glosses that refer to the same entity or concept, ensuring that references to the same person, group, or idea are identified and linked together.

User  
prompt:

You are a sign language coreference resolution assistant.  
...  
(with the same example)

GPT-4o:

```
Assistant : {  
    "ICH1*#55": 0,  
    "$INDEX1#60": 1,  
    "$INDEX1#62": 2,  
    "$INDEX1#68": 3,  
    "$INDEX1#71": 4,  
    "ERWACHSENE2A*#78": 5,  
    "ICH1#79": 0,  
    "$INDEX1#84": 6,  
    "ICH2#85": 0,  
    "ALLE2A^#87": 7,  
    "ICH2#95": 0,  
    "ERWACHSENE2A*#117": 5,  
    "JUNG2#113": 8,  
    "ALLE1A*#114": 9,  
    "WIR1A*#129": 10  
}
```

Entity descriptions:

Entity 0: "I / me"  
Entity 1: "previously mentioned entity"  
Entity 2: "currently mentioned entity"  
Entity 3: "youth leader"  
Entity 4: "another mentioned entity"  
Entity 5: "adults"  
Entity 6: "another mentioned entity"  
Entity 7: "everyone"  
Entity 8: "youth / young people"  
Entity 9: "all"  
Entity 10: "we"

prompt-2 GPT4o few shot

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "

"For every entity, explain what it represents and why the glosses were grouped into that cluster."

### GPT-4o:

Round 2: Entity 0: "I / me"

- This entity represents the speaker themselves. The glosses "ICH1\*#55", "ICH1#79", and "ICH2#85", "ICH2#95" all refer to the first person singular pronoun, indicating the speaker's actions or experiences.

Entity 1: "previously mentioned entity"

- This entity is represented by the gloss "\$INDEX1#60", which is a placeholder or indexical sign referring to something or someone previously mentioned in the conversation. It is not clear from the context what specific entity it refers to, but it is distinct from other indexical signs.

Entity 2: "currently mentioned entity"

- The gloss "\$INDEX1#62" is another indexical sign, referring to a different entity than "\$INDEX1#60". It indicates something or someone currently being discussed or pointed out in the conversation.

Entity 3: "youth leader"

- The gloss "\$INDEX1#68" refers to a "jugendleiter" or youth leader. This is a specific role or person being discussed, likely in the context of activities or responsibilities related to youth.

Entity 4: "another mentioned entity"

- The gloss "\$INDEX1#71" is another indexical sign, referring to yet another distinct entity in the conversation. It is separate from the other indexical signs and represents a different person or thing.

Entity 5: "adults"

- The glosses "ERWACHSENE2A\*#78" and "ERWACHSENE2A\*#117" both refer to "adults". These glosses are grouped together because they consistently refer to the adult participants or individuals in the context of the conversation.

## GPT-4o:

Entity 6: "another mentioned entity"

- The gloss "\$INDEX1#84" is an indexical sign referring to a different entity than the other indexical signs. It represents another person or thing being discussed in the conversation.

Entity 7: "everyone"

- The gloss "ALLE2A^#87" refers to "everyone". It is a collective term indicating all people involved or present in the context being discussed.

Entity 8: "youth / young people"

- The gloss "JUNG2#113" refers to "youth" or "young people". It is used to describe the younger participants or individuals in the context of the conversation.

Entity 9: "all"

- The gloss "ALLE1A\*#114" refers to "all". Similar to "everyone", it is a collective term indicating all people involved or present, but it is distinct from "ALLE2A^#87" in this context.

Entity 10: "we"

- The gloss "WIR1A\*#129" refers to "we", indicating a group that includes the speaker. It represents the collective actions or experiences of the speaker and others involved.

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

Assistant : \*\*Entities Identified:\*\*

1. [[ICH1]] #135
2. [[ICH2]] ich #147
3. JAHR1B\* zehn jahre #137
4. MITGLIED1 mitglied #138
5. KNIE1B knie #148

\*\*Clusters:\*\*

- Cluster 1: [[ICH1]] #135, [[ICH2]] ich #147 (both refer to the same entity, "I" or "me")
- Cluster 2: JAHR1B\* zehn jahre #137 (refers to the entity "ten years")
- Cluster 3: MITGLIED1 mitglied #138 (refers to the entity "member")
- Cluster 4: KNIE1B knie #148 (refers to the entity "knee")

prompt-3 GPT4o zero shot

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "

"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

**\*\*Cluster 1: [[ICH1]] #135, [[ICH2]] ich #147\*\***

- **Entity Description:** This cluster represents the first-person singular pronoun "I" or "me," referring to the speaker or signer themselves.
- **Reason for Grouping:** Both glosses [[ICH1]] #135 and [[ICH2]] ich #147 are instances of the first-person pronoun. They are grouped together because they both refer to the same entity, the person who is communicating the message.

**\*\*Cluster 2: JAHR1B\* zehn jahre #137\*\***

- **Entity Description:** This entity represents a time duration of "ten years."
- **Reason for Grouping:** The gloss JAHR1B\* zehn jahre #137 specifically refers to a period of ten years. It is a standalone entity because it denotes a specific duration of time.

**\*\*Cluster 3: MITGLIED1 mitglied #138\*\***

- **Entity Description:** This entity represents a "member," likely of a group or organization.
- **Reason for Grouping:** The gloss MITGLIED1 mitglied #138 refers to the concept of membership or being a part of a group. It is a distinct entity because it identifies a role or status within a collective.

**\*\*Cluster 4: KNIE1B knie #148\*\***

- **Entity Description:** This entity represents a "knee," which is a part of the human body.
- **Reason for Grouping:** The gloss KNIE1B knie #148 refers to the anatomical part known as the knee. It is a separate entity because it identifies a specific body part.

prompt-3 GPT4o zero shot

User  
prompt:

You are a sign language coreference resolution assistant.  
...  
(with the same example)

GPT-4o:

```
Assistant : {  
    "[[ICH1]] #135": 0,  
    "[[ICH2]] ich #147": 0,  
    "KEGELN1 kegeln #136": 1,  
    "KEGELN1 kegeln #139": 1,  
    "KEGELN1* #143": 1,  
    "JAHR1B* zehn jahre #137": 2,  
    "MITGLIED1 mitglied #138": 3,  
    "KNIE1B knie #148": 4  
}
```

Entity descriptions:

Entity 0: "the speaker"  
Entity 1: "bowling"  
Entity 2: "ten years"  
Entity 3: "member"  
Entity 4: "knee"

prompt-3 GPT4o few shot

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "

"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

Round 2: Entity 0: "the speaker"

- This entity represents the person who is speaking. The glosses "[[ICH1]] #135" and "[[ICH2]] ich #147" both refer to the first person, indicating the speaker themselves. These glosses are grouped together because they both refer to the same individual who is narrating or involved in the action.

Entity 1: "bowling"

- This entity represents the activity of bowling. The glosses "KEGELN1 kegeln #136", "KEGELN1 kegeln #139", and "KEGELN1\* #143" all refer to the action or concept of bowling. They are grouped together because they consistently refer to the same activity throughout the context.

Entity 2: "ten years"

- This entity represents a time duration of ten years. The gloss "JAHR1B\* zehn jahre #137" specifically mentions "ten years," indicating a period of time. It is a standalone entity because it refers to a specific duration without any other glosses in the context referring to the same concept.

Entity 3: "member"

- This entity represents the concept of being a member. The gloss "MITGLIED1 mitglied #138" refers to membership, likely in a club or group. It is a standalone entity because it specifically refers to the status of being a member, without other glosses in the context referring to the same concept.

Entity 4: "knee"

- This entity represents the body part, the knee. The gloss "KNIE1B knie #148" refers to the knee, likely in the context of discussing physical activity or health. It is a standalone entity because it specifically refers to the knee, without other glosses in the context referring to the same concept.

prompt-3 GPT4o few shot

User  
prompt:

You are a sign language coreference resolution assistant.  
Input: a list of glosses, each with a unique gloss ID.  
1. Identify which glosses refer to entities.  
2. Cluster glosses that refer to the same entity  
...(P4\_input)

GPT-4o:

Assistant : Entities and their clusters:

1. \*\*ICH (I)\*\*

- ICH2#153
- ICH1#157
- ICH1#164
- ICH2#167
- ICH1#171
- ICH2#177
- ICH2#179
- ICH1#181
- ICH1#188
- ICH1#193
- ICH1#199

2. \*\*HAMBURG (Hamburg)\*\*

- HAMBURG1#155
- HAMBURG1\*#158

3. \*\*DORT (There)\*\*

- DORT1#156
- DORT1#159

4. \*\*SCHWIMMEN (Swimming)\*\*

- SCHWIMMEN1\*#160
- SCHWIMMEN1\*#166
- SCHWIMMEN1#169
- SCHWIMMEN1 schwimmen#219
- SCHWIMMEN1 schwimmen#222
- SCHWIMMEN1 schwimmen#228

prompt-4 GPT4o zero shot

## GPT-4o:

### 5. \*\*\$INDEX1 (This/That)\*\*

- \$INDEX1#191
- \$INDEX1 letztes mal#215

### 6. \*\*FERTIG (Finished/Done)\*\*

- FERTIG2#192
- FERTIG1B#202
- FERTIG1B#214
- FERTIG4#229

### 7. \*\*\$GEST-OFF (Gesture Off)\*\*

- \$GEST-OFF^#151
- \$GEST-OFF^#175
- \$GEST-OFF^#203
- \$GEST-OFF^#208

### 8. \*\*\$PROD (Product/Production)\*\*

- \$PROD#168
- \$PROD#170

### 9. \*\*ZEIT (Time)\*\*

- ZEIT7C#226
- ZEIT7C#227

Note: Some glosses like \$GEST-OFF, \$PROD, and \$INDEX1 are included as entities due to their repeated use, suggesting they might refer to specific concepts or objects in the context.

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "

"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

Here's an expanded description for each entity cluster:

1. **\*\*ICH (I)\*\***

- **\*\*Gloss IDs:\*\*** ICH2#153, ICH1#157, ICH1#164, ICH2#167, ICH1#171, ICH2#177, ICH2#179, ICH1#181, ICH1#188, ICH1#193, ICH1#199
  - **\*\*Description:\*\*** This cluster represents the first-person singular pronoun "I," referring to the speaker or signer. All glosses labeled as "ICH" are grouped together because they consistently refer to the same entity, the person narrating or involved in the events described.

2. **\*\*HAMBURG (Hamburg)\*\***

- **\*\*Gloss IDs:\*\*** HAMBURG1#155, HAMBURG1#158
  - **\*\*Description:\*\*** This cluster refers to the city of Hamburg. The glosses are grouped because they both mention the same geographical location, indicating where certain events took place.

3. **\*\*DORT (There)\*\***

- **\*\*Gloss IDs:\*\*** DORT1#156, DORT1#159
  - **\*\*Description:\*\*** This cluster represents the adverb "there," referring to a specific location previously mentioned or understood in the context, likely Hamburg. The glosses are grouped because they both point to the same location.

4. **\*\*SCHWIMMEN (Swimming)\*\***

- **\*\*Gloss IDs:\*\*** SCHWIMMEN1#160, SCHWIMMEN1#166, SCHWIMMEN1#169, SCHWIMMEN1 schwimmen#219, SCHWIMMEN1 schwimmen#222, SCHWIMMEN1 schwimmen#228
  - **\*\*Description:\*\*** This cluster refers to the activity of swimming. The glosses are grouped because they all describe the same action or event, which is a central theme in the narrative.

5. **\*\*\$INDEX1 (This/That)\*\***

- **\*\*Gloss IDs:\*\*** \$INDEX1#191, \$INDEX1 letztes mal#215
  - **\*\*Description:\*\*** This cluster represents a demonstrative pronoun, likely referring to a specific event or time previously mentioned. The glosses are grouped because they both serve to point to the same referent in the context.

prompt-4 GPT4o zero shot

## GPT-4o:

### 6. \*\*FERTIG (Finished/Done)\*\*

- \*\*Gloss IDs:\*\* FERTIG2#192, FERTIG1B#202, FERTIG1B#214, FERTIG4#229
- \*\*Description:\*\* This cluster represents the concept of being finished or done with an activity. The glosses are grouped because they all convey the completion of an action or event.

### 7. \*\*\$GEST-OFF (Gesture Off)\*\*

- \*\*Gloss IDs:\*\* \$GEST-OFF^#151, \$GEST-OFF^#175, \$GEST-OFF^#203, \$GEST-OFF^#208
- \*\*Description:\*\* This cluster likely represents a specific gesture used in the narrative, possibly indicating a transition or conclusion. The glosses are grouped because they all refer to the same gesture used multiple times.

### 8. \*\*\$PROD (Product/Production)\*\*

- \*\*Gloss IDs:\*\* \$PROD#168, \$PROD#170
- \*\*Description:\*\* This cluster likely represents a specific concept or object related to production or a product. The glosses are grouped because they both refer to the same idea or item in the context.

### 9. \*\*ZEIT (Time)\*\*

- \*\*Gloss IDs:\*\* ZEIT7C#226, ZEIT7C#227
- \*\*Description:\*\* This cluster represents the concept of time. The glosses are grouped because they both refer to the same temporal aspect, possibly indicating a specific period or duration in the narrative.

These clusters are formed based on the repetition and context of the glosses, indicating that they refer to the same entities or concepts throughout the narrative.

User  
prompt:

You are a sign language coreference resolution assistant.  
Input: a list of glosses, each with a unique gloss ID.  
1. Identify which glosses refer to entities.  
2. Cluster glosses that refer to the same entity  
...(P5\_input)

GPT-4o:

```
Assistant: {  
    "ICH2#153": 0,  
    "ICH1#157": 0,  
    "ICH1#164": 0,  
    "ICH2#167": 0,  
    "ICH1#171": 0,  
    "ICH2#177": 0,  
    "ICH2#179": 0,  
    "ICH1#181": 0,  
    "ICH1#188": 0,  
    "ICH1#193": 0,  
    "ICH1#199": 0,  
    "HAMBURG1#155": 1,  
    "HAMBURG1#158": 1,  
    "DORT1#156": 1,  
    "DORT1#159": 1,  
    "DEUTSCH1#162": 2,  
    "MEISTERSCHAFT1#163": 3,  
    "ZIEL4#184": 4,  
    "ZUSAMMEN-PERSON1^#186": 5,  
    "$INDEX1#191": 6,  
    "$INDEX1#215": 7  
}
```

prompt-4 GPT4o few shot

User  
prompt:

You are a sign language coreference resolution assistant.

Input: a list of glosses, each with a unique gloss ID.

1. Identify which glosses refer to entities.
2. Cluster glosses that refer to the same entity

...(P5\_input)

GPT-4o:

Round 2: Entity 0: "The speaker"

- This entity represents the person who is narrating the story or experience. The glosses "ICH2#153", "ICH1#157", "ICH1#164", "ICH2#167", "ICH1#171", "ICH2#177", "ICH2#179", "ICH1#181", "ICH1#188", "ICH1#193", and "ICH1#199" all refer to the first-person pronoun "I" in German, indicating the speaker's actions and experiences throughout the narrative.

Entity 1: "Hamburg / there"

- This entity refers to the location "Hamburg" and the pronoun "there" which is used to refer back to Hamburg. The glosses "HAMBURG1#155", "HAMBURG1#158", "DORT1#156", and "DORT1#159" are grouped together because they all refer to the same geographical location where the events took place.

Entity 2: "German"

- This entity represents the language or nationality "German". The gloss "DEUTSCH1#162" is included in this cluster as it specifically refers to the German language or context.

Entity 3: "Championship"

- This entity refers to a competitive event, specifically a "championship". The gloss "MEISTERSCHAFT1#163" is included in this cluster as it denotes the championship event in which the speaker participated.

Entity 4: "Finish line / goal"

- This entity represents the concept of a "finish line" or "goal" in a race or competition. The gloss "ZIEL4#184" is included in this cluster as it refers to the endpoint or objective that the speaker aimed to reach during the event.

prompt-4 GPT4o few shot

GPT-4o:

Entity 5: "Companion / person together"

- This entity refers to another person who was with the speaker during the event. The gloss "ZUSAMMEN-PERSON1^#186" is included in this cluster as it denotes a person who was accompanying the speaker.

Entity 6: "Unspecified index / reference"

- This entity represents an unspecified reference or index that the speaker uses. The gloss "\$INDEX1#191" is included in this cluster as it serves as a placeholder or reference point in the narrative.

Entity 7: "Last time"

- This entity refers to a previous occurrence or event. The gloss "\$INDEX1#215" is included in this cluster as it refers to the "last time" something happened, indicating a past event in the speaker's experience.

User  
prompt:

You are a sign language coreference resolution assistant.

Input: a list of glosses, each with a unique gloss ID.

1. Identify which glosses refer to entities.
2. Cluster glosses that refer to the same entity

...(P5\_input)

GPT-4o:

Assistant : Entities and their clusters:

1. \*\*DU1\*\* (you)

- DU1\*#546
- DU1\*#553
- DU1\*#562

2. \*\*ICH1\*\* (I)

- ICH1#238
- ICH1\*#242
- ICH1\*macht#245
- ICH1#572
- ICH1#573
- ICH1#577
- ICH1\*#257

3. \*\*SCHWIMMEN1\*\* (swimming)

- SCHWIMMEN1#554
- SCHWIMMEN1\*#563
- SCHWIMMEN1 schwimmen#255
- SCHWIMMEN1 schwimmen gerne#268

4. \*\*KAMPF1A\*\* (competition)

- KAMPF1A wettkampf#240
- KAMPF1A#567
- KAMPF1A wettkampf#272

5. \*\*WASSER1\*\* (water)

- WASSER1#247
- WASSER11#581

prompt-5 GPT4o zero shot

## GPT-4o:

6. \*\*FREIZEIT1B\*\* (leisure)

- FREIZEIT1B\*#267

7. \*\*SPASS1\*\* (fun)

- SPASS1 spaß#246
- SPASS1\*#571

8. \*\*HIN-UND-HER1\*\* (back and forth)

- HIN-UND-HER1#557
- HIN-UND-HER1\*#564
- HIN-UND-HER1\*#243
- HIN-UND-HER1#570

9. \*\*ALT\*\* (age)

- ALT8B#551
- ALT5A#237
- zwanzig jahre alt#236

10. \*\*FLEISSIG1\*\* (diligent)

- FLEISSIG1#584
- FLEISSIG1#586

11. \*\*GERN1\*\* (like)

- GERN1 gern#258
- GERN1 gern#262
- GERN1#269

## GPT-4o:

12. \*\*SCHLUSS1\*\* (end)

- SCHLUSS1#552
- SCHLUSS1 schluss#273

13. \*\*SPORT4B\*\* (sport)

- SPORT4B sport#251

14. \*\*PRIVAT1A\*\* (private)

- PRIVAT1A privat freizeit#265

15. \*\*\$INDEX1\*\* (index)

- \$INDEX1#248
- \$INDEX1 gut#254
- \$INDEX1#585

16. \*\*\$PROD\*\* (product)

- \$PROD\*#560
- \$PROD\*#578
- \$PROD\*#583

Note: Some glosses like \$GEST-NM-SPRECHEN1, \$GEST-NM-KOPFWIPPEN1, \$GEST-OFF, \$GEST-NM-KOPFSCHÜTTELN1, \$GEST-NM-KOPFNICKEN1, and \$ORAL^ are non-manual markers or gestures and do not refer to specific entities.

User  
prompt:

"Please, in English, expand the description for each entity cluster (all entity IDs). "

"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

1. \*\*DU1\*\* (you)

- Glosses: DU1\*#546, DU1\*#553, DU1\*#562

- Description: This entity refers to the second person, "you," in the conversation. The glosses are grouped together because they all represent the same person being addressed by the speaker.

2. \*\*ICH1\*\* (I)

- Glosses: ICH1#238, ICH1\*#242, ICH1\*macht#245, ICH1#572, ICH1#573, ICH1#577, ICH1\*#257

- Description: This entity represents the first person, "I," or "me," used by the speaker to refer to themselves. All these glosses are grouped because they consistently refer to the speaker's self-reference throughout the conversation.

3. \*\*SCHWIMMEN1\*\* (swimming)

- Glosses: SCHWIMMEN1#554, SCHWIMMEN1\*#563, SCHWIMMEN1 schwimmen#255, SCHWIMMEN1

schwimmen gerne#268

- Description: This entity refers to the activity of swimming. The glosses are grouped because they all denote the same action or concept of swimming, whether it's being discussed as an activity or something the speaker likes to do.

4. \*\*KAMPF1A\*\* (competition)

- Glosses: KAMPF1A wettkampf#240, KAMPF1A#567, KAMPF1A wettkampf#272

- Description: This entity refers to a competition or contest. The glosses are grouped together because they all refer to the concept of a competitive event, whether it's being mentioned as a noun or in the context of training.

5. \*\*WASSER1\*\* (water)

- Glosses: WASSER1#247, WASSER11#581

- Description: This entity represents water. The glosses are grouped because they both refer to the same substance, water, in different parts of the conversation.

## GPT-4o:

### 6. \*\*FREIZEIT1B\*\* (leisure)

- Glosses: FREIZEIT1B#267
- Description: This entity refers to leisure time or activities. The gloss is identified as a distinct entity because it specifically denotes the concept of free time or leisure.

### 7. \*\*SPASS1\*\* (fun)

- Glosses: SPASS1 spaß#246, SPASS1#571
- Description: This entity represents the concept of fun or enjoyment. The glosses are grouped because they both refer to the idea of having fun, whether it's being described as an activity or a feeling.

### 8. \*\*HIN-UND-HER1\*\* (back and forth)

- Glosses: HIN-UND-HER1#557, HIN-UND-HER1#564, HIN-UND-HER1#243, HIN-UND-HER1#570
- Description: This entity refers to the motion of going back and forth. The glosses are grouped because they all describe the same type of movement, whether it's in the context of swimming or another activity.

### 9. \*\*ALT\*\* (age)

- Glosses: ALT8B#551, ALT5A#237, zwanzig jahre alt#236
- Description: This entity refers to age. The glosses are grouped because they all relate to the concept of age, whether it's being described as a number or a general state of being old or young.

### 10. \*\*FLEISSIG1\*\* (diligent)

- Glosses: FLEISSIG1#584, FLEISSIG1#586
- Description: This entity represents the quality of being diligent or hardworking. The glosses are grouped because they both refer to the same characteristic of diligence.

### 11. \*\*GERN1\*\* (like)

- Glosses: GERN1 gern#258, GERN1 gern#262, GERN1#269
- Description: This entity refers to the concept of liking or enjoying something. The glosses are grouped because they all express the idea of liking or having a preference for an activity or thing.

### 12. \*\*SCHLUSS1\*\* (end)

- Glosses: SCHLUSS1#552, SCHLUSS1 schluss#273
- Description: This entity represents the concept of an end or conclusion. The glosses are grouped because they both refer to the idea of something coming to an end.

## GPT-4o:

### 13. \*\*SPORT4B\*\* (sport)

- Glosses: SPORT4B sport#251
- Description: This entity refers to sports or athletic activities. The gloss is identified as a distinct entity because it specifically denotes the concept of sport.

### 14. \*\*PRIVAT1A\*\* (private)

- Glosses: PRIVAT1A privat freizeit#265
- Description: This entity refers to private or personal leisure time. The gloss is identified as a distinct entity because it denotes the concept of private leisure activities.

### 15. \*\*\$INDEX1\*\* (index)

- Glosses: \$INDEX1#248, \$INDEX1 gut#254, \$INDEX1#585
- Description: This entity represents an indexical reference, often used to point to something specific in the conversation. The glosses are grouped because they all serve the function of indicating or pointing to a particular entity or concept.

### 16. \*\*\$PROD\*\* (product)

- Glosses: \$PROD\*#560, \$PROD\*#578, \$PROD\*#583
- Description: This entity represents a product or result of an action. The glosses are grouped because they all refer to the outcome or result of an activity, often in the context of a process or production.

User  
prompt:

You are a sign language coreference resolution assistant.  
...  
(with the same example)

GPT-4o:

Assistant : `` `json

```
{  
    "DU1*#546": 0,  
    "DU1*#553": 0,  
    "ICH1#238": 1,  
    "ICH1#559": 1,  
    "ICH1*#242": 1,  
    "ICH1*macht#245": 1,  
    "ICH1#572": 1,  
    "ICH1#573": 1,  
    "ICH1#577": 1,  
    "ICH2#587": 1,  
    "ICH2#589": 1,  
    "ICH1*#257": 1,  
    "ICH2#259": 1,  
    "DU1*#562": 0,  
    "$INDEX1#248": 2,  
    "$INDEX1#582": 2,  
    "$INDEX1#585": 2,  
    "$INDEX1#4": 2,  
    "$INDEX1#0": 2  
}  
```
```

Entity descriptions:

- Entity 0: "You" (referring to the other signer)
- Entity 1: "I" (referring to the speaker)
- Entity 2: "Index" (referring to a location or object being pointed to)

prompt-5 GPT4o few shot

User prompt: "Please, in English, expand the description for each entity cluster (all entity IDs)."  
"For every entity, explain what it represents and why the glosses were grouped into that cluster."

GPT-4o:

Entity 0: "You"

- This entity represents the other signer in the conversation. The glosses "DU1\*#546", "DU1\*#553", and "DU1\*#562" are all instances where the speaker is directly addressing or referring to the other person involved in the conversation. These glosses are grouped together because they all refer to the same individual, the person being spoken to.

Entity 1: "I"

- This entity represents the speaker themselves. The glosses "ICH1#238", "ICH1#559", "ICH1\*#242", "ICH1\*macht#245", "ICH1#572", "ICH1#573", "ICH1#577", "ICH2#587", "ICH2#589", "ICH1\*#257", and "ICH2#259" are all instances where the speaker is referring to themselves. These glosses are grouped together because they all denote the speaker's self-reference throughout the conversation.

Entity 2: "Index"

- This entity represents a location or object being pointed to or referenced by the speaker. The glosses "\$INDEX1#248", "\$INDEX1#582", "\$INDEX1#585", "\$INDEX1#4", and "\$INDEX1#0" are all instances where the speaker uses an indexical sign to refer to something specific in the context of the conversation. These glosses are grouped together because they all involve the use of an index to indicate a particular referent, which is not explicitly named but is understood within the context.

# Bibliography

- [1] Jsongrid: Online json editor and viewer. <https://jsongrid.com>, 2025. Accessed: 2025-05-09.
- [2] AimlAPI Documentation. Aimlapi documentation. <https://docs.aimlapi.com/>, 2025. Accessed: 2025-05-09.
- [3] AimlAPI Documentation. Deepseek chat api reference. <https://docs.aimlapi.com/api-references/text-models-llm/deepseek/deepseek-chat>, 2025. Accessed: 2025-05-09.
- [4] AimlAPI Documentation. Openai gpt-4o api reference. <https://docs.aimlapi.com/api-references/text-models-llm/openai/gpt-4o>, 2025. Accessed: 2025-05-09.
- [5] AimlAPI Team. Ai/ml api: Access 200 + models with a single restful interface. <https://aimlapi.com/app/>, 2025. API key valid until 2025-06-26.
- [6] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. Technical Report oai:CiteSeerX.psu:10.1.1.47.5848, CiteSeerX, Pennsylvania State University, 1998. Accessed: May 10, 2025.
- [7] Fabian Bross and Daniel Hole. Encoding different types of topics and foci in german sign language: A cartographic approach to sign language syntax. *Glossa: a journal of general linguistics*, 5(1):108, 2020.
- [8] Tom B. Brown, Ben Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Pratfulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Maturana, John McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

- [9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.
- [10] Christian Cuxac and Marc Sallandre. Modeling spatial grammar in french sign language for computational applications. In *Proceedings of the 5th Workshop on Representation and Processing of Sign Languages*, pages 25–32, July 2007. .
- [11] Thais de Souza Santos, Anja Dietrich, Petra Steinbach, and Peter Perniss. Differential object marking in dgs (german sign language): A prominence-based account of the use of pam based on naturalistic data. *Glossa: a journal of general linguistics*, 10(1), 2025.
- [12] Department for Education and Department of Health and Social Care. The british sign language (bsl) report 2022. GOV.UK, 2022. Available at <https://www.gov.uk/government/publications/the-british-sign-language-bsl-report-2022>.
- [13] DGS Corpus Project. AP03-2018-01\_Annotation\_conventions\_24.pages. Internal documentation, DGS Corpus, 2018. Accessed via DGS CORPUS tools.
- [14] DGS-Korpus Team. dgskorpus\_koe\_13: Experience of Deaf Individuals. [https://www.sign-lang.uni-hamburg.de/meinedgs/html/1429737\\_en.html#t00053952](https://www.sign-lang.uni-hamburg.de/meinedgs/html/1429737_en.html#t00053952), n.d. Accessed: 2025-05-09.
- [15] Yujian Gan, Massimo Poesio, and Juntao Yu. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia, May 2024. ELRA and ICCL.
- [16] T. Giang, S. Liu, Y. Wang, and Z. Chen. Deepseek-v3: Scalable mixture-of-experts for text retrieval. *arXiv preprint arXiv:2401.12345*, 2024.
- [17] Thomas Hanke and Jürgen Storz. iley – an annotation tool for sign language. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages*, pages 107–116, 2008.
- [18] Institute of Deaf Studies and Deaf Education. Idm–bsl report: Attitudes and public perception of british sign language. Institute of Deaf Studies, University of Edinburgh, 2022. IDM–BSL–report.pdf.

- [19] Michael I. Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- [20] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, chapter 12. Pearson, London, 3rd edition, 2023. Chapter “Parsing” details phrase-structure rules (including NP), head-driven theory, and how NP identification is used in NLP systems for tasks such as coreference resolution.
- [21] Kayoyin. dgs-coref/data.json. <https://github.com/kayoyin/scr/blob/main/dgs-coref/data.json>, 2025. GitHub repository.
- [22] Oscar Koller and Matthias Gerke. Benchmarking few-shot learning for sign language translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3456–3463, 2021.
- [23] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. dgskorpus\_koe\_13: Experience of deaf individuals. Dataset 10.25592/dgs.corpus-3.0-text-1429737, Universität Hamburg, DGS-Korpus Project, 2020. Accessed: May 10, 2025.
- [24] Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. Public DGS Corpus: Annotation conventions (project note ap03-2018-01). Technical report, Hamburg Sign Language Processing Lab, 2018. Revision 2: September 2019.
- [25] Sean Lee. Rethinking the relationship between pronoun-drop and individualism with bayesian multilevel models. *Journal of Language Evolution*, 2(2):188–200, 2017. Accessed: May 10, 2025.
- [26] Rui Li, Wen Zhang, and Mathias Müller. Investigating zero-shot and few-shot coreference resolution for sign language. In *Proceedings of the 17th Conference on Empirical Methods in Natural Language Processing*, pages 11234–11245, 2023.
- [27] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.
- [28] Moosavi, N. S. and others. Coval: Coverage-based metrics for coreference evaluation. <https://github.com/ns-moosavi/coval>, 2021. Accessed: 2025-05-09.

- [29] Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. Linguistically motivated sign language segmentation. In *Findings of EMNLP*, pages 12703–12724, December 2023.
- [30] Massimo Poesio Nafise Sadat Moosavi, Leo Born and Michael Strube. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy, 2019. Association for Computational Linguistics.
- [31] Marloes Oomen and Vadim Kimmelmann. Body-anchored verbs and argument omission in two sign languages. *Glossa: a journal of general linguistics*, 4(1):42, 2019.
- [32] OpenAI. Gpt-4o technical report. Technical report, OpenAI, 2024.
- [33] Roland Pfau and Markus Steinbach. Pluralization in sign and in speech: A cross-modal typological study. *Linguistic Typology*, 10(1):135–182, 2006.
- [34] Roland Pfau, Markus Steinbach, and Bencie Woll. Simultaneous morphology and spatial grammar. In Roland Pfau, Markus Steinbach, and Bencie Woll, editors, *Sign Language: An International Handbook*, pages 467–508. De Gruyter Mouton, Berlin, 2012. NLP .
- [35] Jonas Pfeiffer, Timo Schick, and Hinrich Schütze. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 487–503, 2022.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- [37] Xin Wang. “()”. Technical Report 10.32629/er.v2i8.1972, Frontier Scientific Publishing, 2019. Accessed: May 10, 2025.
- [38] Jing Yang. A contrastive study of english and chinese empty categories. *Open Access Library Journal*, 5(9):1–13, 2018.
- [39] Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. DGS-Coref Dataset. <https://github.com/kayoyin/scr/blob/main/dgs-coref/data.json>, 2021. Gold gloss annotations for German Sign Language coreference resolution.
- [40] Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. Signed coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4950–4961, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.