# Exploring Coreference Resolution in Glosses of German Sign Language

**Yuzheng Bao** and **Haixia Chai**
Department of Computing Science
School of Natural and Computing Sciences
University of Aberdeen
`{y.bao.21, haixia.chai}@abdn.ac.uk`

## Abstract

In recent years, research on sign languages has attracted increasing attention in the NLP community and requires more effort from a linguistic perspective. In this paper, we explore coreference resolution in German Sign Language (GSL) primarily through gloss-based analysis. Specifically, in GSL glosses, we conduct a linguistic analysis of coreference, add coreference annotations based on three videos, and evaluate the ability of two large language models to resolve coreference. We gain valuable insights into coreference resolution in GSL, which pave the way for future research.[1]

## 1 Introduction

Natural language develops naturally for daily communication among humans. As a first language for deaf and hearing impaired individuals, sign languages (e.g., American Sign Language and German Sign Language) are visual-spatial natural languages with their own sophisticated linguistic systems—including lexicon, morphology, phonology, syntax, and pragmatics—not only gestures (Stokoe, 1980). The complexity of signs lies in the fact that they can be articulated through multiple phonological units, including handshape, palm orientation, position, and facial expressions (e.g., eyebrow movement and head motion) within a three-dimensional signing space (Herrmann and Steinbach, 2011; Michael et al., 2011). Compared to spoken language, sign language remains underexplored and demands linguistic insights from the Natural Language Processing community (Yin et al., 2021b).

| |
|---|
| `GEBÄRDEN1A LEHRER5 ICH1 BIS-HEUTE2 $GEST-OFF^*` |
| *(To this day, I am still a sign language teacher.)* |

Table 1: An example text shows the glosses and their English translation. Glosses are written in capital letters and may include affixes or markers indicating additional information.

To record and analyse signs in written form, glosses annotated by experts are used as linear labels that approximate the semantic meaning of each sign, typically using the base form of a corresponding word in spoken language. Table 1 presents an example of glosses used in German Sign Language. While glosses are widely used as an intermediate step in the study of sign language translation, from signs to spoken text (Müller et al., 2023; Fayyazsanavi et al., 2024), and sign language production, from spoken text to signs (Varanasi et al., 2024; Fang et al., 2024), they can also support documentation, education, and linguistic research.

In this paper, we explore coreference—a linguistic phenomenon crucial for natural language understanding—in German Sign Language, primarily through gloss-based analysis. Although glosses—one-dimensional sequences of words—cannot fully capture spatial constructions or represent the meaning of signs through various cues, such as non-manual features (Yin and Read, 2020; Müller et al., 2023), we focus on coreference in glosses as an initial step towards the Signed Coreference Resolution task (Yin et al., 2021a). To our knowledge, Yin et al. (2021a) is the only recent work that addresses coreference resolution in sign language, specifically for pronominal indexing signs. Following the work, we study entity coreference resolution in German Sign Language using the Public DGS Corpus (Hanke et al., 2020). Our contributions are threefold:

- A linguistic analysis of signed coreference in glosses, including noun phrase, pronoun, ellipsis, and others.
- Entity coreference annotations on the glosses of three videos from the DGS corpus.

---

[1] Our annotated data are publicly available at `https://github.com/orcastimulatee/Coref_GLS_GSL.git`

*Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025)*, pages 77–84
Suzhou, China, November 9, 2025. ©2025 Association for Computational Linguistics

- Evaluating the ability of GPT-4o (OpenAI, 2024) and DeepSeek-v3 (DeepSeek-AI, 2025) to perform coreference resolution on glosses with coreference gold annotations using prompt engineering.

## 2 Related Work

### 2.1 Coreference Resolution in Sign Language

The coreference phenomenon in sign languages has previously been studied (Steinbach and Onea, 2015), e.g., in American Sign Language (Kegl, 1987) and in German Sign Language (Wienholz et al., 2018). More recently, Yin et al. (2021a) introduced the Signed Coreference Resolution task for Sign Language Processing, thereby presenting a new challenge for the coreference research community. The work proposes DGS-Coref, a gloss-based dataset derived from the DGS corpus (Hanke et al., 2020) with coreference annotations. Similarly, to reduce the overhead of visual processing, we conduct annotations on glosses as well. The main difference is that Yin et al. (2021a) focuses solely on pronominal coreference relations, whereas our annotations cover all types of entity coreference. The authors (Yin et al., 2021a) also propose a linguistically informed unsupervised coreference resolution model for the task, using both glosses and spatial features extracted from pose estimations.

### 2.2 Large Language Models

Many studies focus on transformer-based large language models (LLMs) for sign language translation and production tasks (Camgoz et al., 2020; Yin and Read, 2020; Fang et al., 2024), aiming to make them accessible to deaf and signing communities. However, the extent to which an LLM truly understands the underlying structure and meaning of signed discourse remains unclear. Gan et al. (2024) examines the coreference resolution abilities of several LLMs (e.g., the GPT series and the LLaMA2 family) in English by using prompts and conducting both manual and automatic analyses. In contrast, our work evaluates LLMs in the context of sign language. Coreference in sign language involves manual features, non-manual features, and spatial referencing, which differ significantly from the devices used in spoken languages.

## 3 Coreference in Sign Language

In sign language, a signer can introduce a new entity into the discourse not only through explicit signs that visually depict its shape, orientation, or movement, but also by assigning it a spatial locus within the signing space in front of the signer, which can later be used to refer back to the entity. To understand how to track an entity, we conduct a basic analysis of the Public DGS Corpus.

**Noun Phrase.** Iconicity is one of the prominent properties of sign language (Perniss et al., 2010). For example, `$PROD SCHWIMMEN` (*swimming*) is a gloss for a productive sign used to represent an inanimate entity—swimming. It illustrates the backstroke, a swimming style, with alternating backward circular motions of both hands. It is worth noting that *backstroke* is not annotated in glosses, but only in the mouthing unit (i.e., it is coarticulated with mouthed German words for lip reading). This means that mentions referring to the same entity can appear in different units, making coreference resolution a task in a multidimensional space. Beyond standard glosses, compound glosses, e.g., `TAUB-GEHÖRLOS` (*deaf*), consist of two or more glosses connected by hyphens to express meaning more efficiently. A sequence of glosses is not annotated as a full German sentence but rather as a chunk of a sentence that conveys a core meaning. Therefore, noun phrases may be omitted from either the subject or object position, regardless of word order.

**Pronoun.** `$INDEX` gloss represents an indexical sign (a pointing gesture) that refers to a spatial locus associated with a lexical sign, thus defining a referent. Subsequent pronouns refer back to the referent by pointing to the same locus. It is clear when only two referents are assigned to the left-hand and right-hand areas, respectively. However, if many referents need to be assigned within a limited signing space, the loci can become very close together, making them difficult to differentiate and potentially causing ambiguity in entity resolution. Adding to the complexity, a signer may relocate a referent to a new locus (Yin et al., 2021a). In the DGS corpus, following the gloss annotation convention (Konrad et al., 2018), a number is added as a suffix to the end of a gloss. For example, `ICH1` (*I1*) and `ICH2` (*I2*) are two contextual variants of `ICH` (*I*), though the basic meaning remains the same. In our annotation presented in §4, we treat these two glosses as referring to the same entity. Unlike in English, pronouns in sign language are not morphologically marked for gender (i.e., there are no separate signs for *he* or *she*). This means that gender must be inferred from the discourse context. We observe that, sometimes, pronouns can be replaced with an iconic sign. For example, a signer may use both hands with palms facing each other, circling slightly from the sides towards the center to express the concept of *together*, visually representing a group of people, rather than signing *us* directly. As a result, the use of pronouns is, to some extent, reduced.

**Ellipsis.** Ellipsis is a common phenomenon in sign languages, including but not limited to German Sign Language. It arises from features such as spatial referencing and role shift—a linguistic device that marks different characters through body shift, eye gaze, and head orientation (Proske et al., 2020)—allowing signers to omit overt pronouns or explicit entity names. Moreover, many sign languages are topic-prominent, e.g., WASSER1 $INDEX1 FRISCH1 IMMER4A* (*Yes, you always feel refreshed when you are in the water.*), meaning that a topic or entity (i.e., *water*) is typically introduced first in discourse. This entity may later be omitted from subject or object positions, if it is predictable and inferable from context. In Table 2, we present another example of ellipsis, in which even more expressions are not explicitly signed. Signer A was talking about playing ninepin bowling in the earlier context. Signer B inferred that the reason for the knee pain was kneeling down too much, so only the corresponding sign, HINKNIEN-SICH1 (*kneel-down*), was used, omitting other non-essential signs. Resolving and recovering such elliptical constituents can undoubtedly benefit sign language understanding and gloss-based studies of sign language translation. In spoken language, many linguistic theories have been studied in relation to zero pronouns and focus, including topic chain theory (Tsao, 1977; Zhang et al., 2022) and centering theory (Joshi and Weinstein, 1981; Grosz et al., 1983, 1995; Walker et al., 1998; Chai and Strube, 2022). We raise a question of whether these theories could aid in resolving implicit expressions in sign language, which we leave for future work.

| | |
|---|---|
| A: | JETZT1* ICH2 KNIE1A* SCHMERZ3 $GEST-OFF^* RÜCKEN-UNTEN1E SCHMERZ3<br>*(Now I have knee and back pain.)* |
| B: | HINKNIEN-SICH1<br>*(Well, because you had to kneel down a lot.)* |

Table 2: An example illustrating ellipsis in a dialogue between two signers.

**Others.** In the signing space, verb inflection—through modified movements that match the loci of the subject and object to indicate agreement—can help track entities. However, glosses are written in their base form, and verbs are not morphologically inflected in the gloss itself. To this end, cues from visual processing become especially important for discourse understanding and entity resolution.

## 4 Human Annotation

To examine the current ability of LLMs to resolve coreference in sign language glosses (see §5), we conducted coreference annotation on the glosses of three videos totaling 990 seconds from the DGS corpus[2][3][4]. Some gloss names are followed by numerical or alphabetical suffixes to distinguish lexical and phonological variants (Konrad et al., 2018). These glosses were annotated carefully by considering the context, the English translation, and by watching the video. Entities are annotated throughout the entire duration of each video, and singletons are excluded from the annotation. We have two annotators with backgrounds in computer science and computational linguistics, both of whom have knowledge of German Sign Language. Inter-annotator agreement was measured using Krippendorff's $\alpha$ (Krippendorff, 1980), resulting in a high score of 0.93. For the annotations with disagreements, the annotators discussed and reached final decisions for the gold annotations. Ambiguous cases were excluded. We release the annotated data as a JSON file.[5] Table 3 shows the statistics of our annotations.

| | Video1 | Video2 | Video3 |
|---|---|---|---|
| #mentions | 166 | 80 | 90 |
| #noun phrases | 78 | 51 | 42 |
| #pronouns | 63 | 9 | 29 |
| #$INDEX | 25 | 20 | 19 |
| #entities | 36 | 26 | 27 |
| #mentions/entities | 4.6 | 3.0 | 3.3 |

Table 3: Statistics of the annotated data. Video1, Video2, and Video3 refer respectively to the links in the footage[2][3][4].

---

[2] https://www.sign-lang.uni-hamburg.de/meinedgs/html/1429737_en.html

[3] https://www.sign-lang.uni-hamburg.de/meinedgs/html/1183720-17021701-17054739_en.html

[4] https://www.sign-lang.uni-hamburg.de/meinedgs/html/1182135_en.html

[5] Our annotated data are publicly available at https://github.com/orcastimulatee/Coref_GLS_GSL.git

|  | MUC |  |  | $B^3$ |  |  | $CEAF_e$ |  |  | CoNLL |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | R | P | F1 | R | P | F1 | R | P | F1 | R | P | F1 |
| DS-v3_zs | 67.60 | 64.82 | 66.10 | 64.65 | 59.58 | 61.96 | 67.23 | 53.83 | 59.80 | 66.49 | 59.41 | 62.62 |
| DS-v3_fs | **76.20** | 54.58 | 63.60 | **72.79** | 50.65 | 59.70 | 71.39 | 53.50 | 61.18 | **73.46** | 52.91 | 61.49 |
| GPT-4o_zs | 64.30 | 70.85 | 67.38 | 58.81 | 66.20 | 62.26 | 65.39 | 52.09 | 57.98 | 62.83 | 63.05 | 62.54 |
| GPT-4o_fs | 74.50 | **70.95** | **72.70** | 71.68 | **67.10** | **69.30** | **73.42** | **57.54** | **64.50** | 73.20 | **65.20** | **68.83** |

Table 4: Performance on the annotated data in §4. *zs* and *fs* denote the zero-shot and few-shot settings for the two LLMs: DeepSeek-v3 (DS) and GPT-4o. Bold numbers indicate the highest score in each column.

## 5 Prompt Engineering

We employ prompt engineering in zero-shot and few-shot settings (Brown et al., 2020; Liu et al., 2023) to evaluate two LLMs: GPT-4o (OpenAI, 2024) and DeepSeek-v3 (DeepSeek-AI, 2025). Figure 1 shows the prompt template, which includes an instruction, input glosses, and one example in the few-shot setting. For few-shot prompting, we select examples that are similar to the cases in the input glosses under examination to provide the LLMs with additional cues for coreference resolution. Additionally, LLMs are required to provide explanations of their resolved results to enable further manual analysis of their performance. One example prompt is presented in Appendix A.

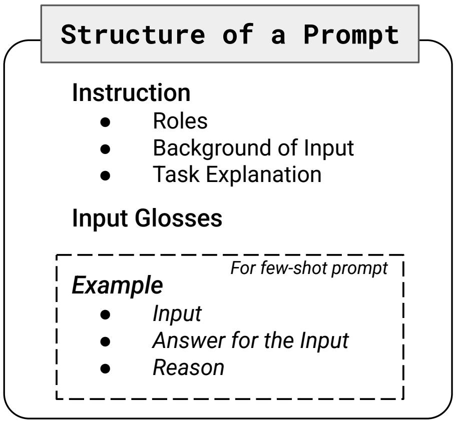


Figure 1: A prompt template shows each components of the prompt. The dashed box applies only to the few-shot setup.

## 6 Experiment

### 6.1 Setup

We conduct experiments on the annotated data (see §4). To obtain stable and reliable responses from LLMs, we divided the data into 20 prompts, grouped by topic for potential further analysis. GPT-4o and DeepSeek-v3 are prompted with a temperature of 0 and a maximum token limit of 5,025 via an AI model API platform[6]. We report macro-averaged results for the 20 prompts in both zero-shot and few-shot settings for the two LLMs. The evaluation uses the CoNLL F1 score, which averages MUC (Vilain et al., 1995), $B^3$ (Bagga and Baldwin, 1998), and CEAFe (Luo, 2005).

### 6.2 Results and Analyses

Table 4 shows that *GPT-4o_fs* achieves the best F1 score of 68.83, which is more than 7 points higher than *DS-v3_fs* and 6 points higher than *GPT-4o_zs*. This indicates that providing an example containing gold coreference annotations improves *GPT-4o*'s overall performance. We also observe that *DS-v3_fs* has high recall (73.46) but low precision (52.91), resulting in a lower CoNLL F1 score. This suggests that *DS-v3_fs* resolves many entities, but few of them are correct. Overall, the two LLMs demonstrate moderate performance on our annotated data. It is important to note that the experiments are conducted on glosses composed of words from spoken language, which are used for training the LLMs. How well they can perform directly on signs remains a worthwhile question for future research.

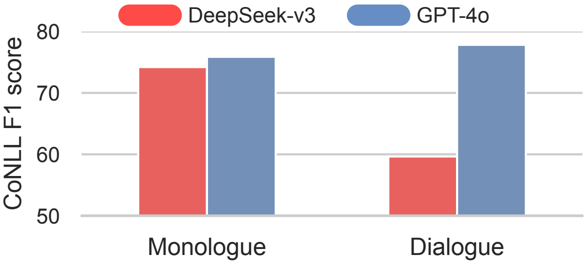


Figure 2: CoNLL F1 scores of the two LLMs in the few-shot setting across different genres.

Figure 2 presents the performance of the two LLMs on the first video, comprising 8 monologue prompts and 5 dialogue prompts. It shows that *GPT-4o_fs* performs better in dialogue than in monologue. Based on our manual analysis of the model

---

[6] `https://aimlapi.com/`

responses, we found that *GPT-4o* can understand the conversation well and can resolve *ICH* (*I*) from Signer A and *DU* (*you*) from Signer B as referring to the same entity. However, *DS-v3_fs* struggles to resolve some entities in dialogue, even when provided with an example and its explanation in the few-shot setting.

We also perform an analysis of some entities that are not resolved successfully. Specifically, for the entity in Table 5, we observe that *DS-v3_fs* can capture the hint from the selected example that is similar to the queried input and correctly resolve the entity. However, *GPT-4o_fs* ignores the example despite our various attempts and fails to resolve it. This suggests that, in this specific case, *GPT-4o_fs* relies more on its internal knowledge and reasoning, making it less influenced by the provided examples, whereas *DS-v3_fs* appears more receptive to such guidance.

| | |
|---|---|
| A: | `$LIST1:1of1d KEGELN1 $LIST1:2of2d SCHERE1* $LIST1:3of3d $NUM-EINER1A:3d BAHN-WEG1A* $GEST-OFF^* $LIST1:4of4d ASPHALT1*`<br>*(For example Bohle, Schere, three lane alleys and classic.)* |
| B: | `$ORAL^KANN1 ALLES1A`<br>*(Can you play all four disciplines?)* |

Table 5: A snippet of glosses between two signers with coreference annotations shown in red.

## 7 Discussions and Conclusions

While analysing the properties of coreference in sign language, our study is primarily based on linear glosses. Further research is needed in coreference resolution modeling—especially for understanding spatial relations—or in developing enhanced glosses that recover omitted elements to support downstream tasks. In this paper, as an initial step, we gain linguistic insights into coreference, annotate glosses accordingly, and evaluate the coreference resolution abilities of two LLMs, to foster future advancements.

## Limitations

Our study is conducted on written glosses, which may omit some information (e.g., from mouthings or productive signs), and therefore the experiments on coreference resolution may not fully reflect a natural signing scenario. Due to limited resources, no deaf people or sign language users were involved in the annotation process.

## Acknowledgments

We thank the anonymous reviewers for their thoughtful comments and constructive feedback, which greatly improved the clarity and quality of this work.

## References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation*, pages 563–566, Granada, Spain.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Haixia Chai and Michael Strube. 2022. Incorporating centering theory into neural coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2996–3002, Seattle, United States. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. 2024. Signllm: Sign language production large language models. *arXiv preprint arXiv:2405.10718*.

Pooya Fayyazsanavi, Antonios Anastasopoulos, and Jana Kosecka. 2024. Gloss2Text: Sign language gloss translation using LLMs and semantically aware label smoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16162–16171, Miami, Florida, USA. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. 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. ELRA and ICCL.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

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

Annika Herrmann and Markus Steinbach. 2011. Nonmanuals in sign languages. *Sign Language & Linguistics*, 14(1):3–8. Publisher: John Benjamins.

Aravind K. Joshi and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. In *Proceedings of the IJCAI, Vancouver, CA*, pages 385–387.

Judy Kegl. 1987. *Coreference Relations in American Sign Language*, pages 135–170. Springer Netherlands, Dordrecht.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2018. Public DGS Corpus: Annotation Conventions. *Technical report, Project Note AP03–2018-01, DGS-Korpus project*.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. CA: Sage Publications, Beverly Hills.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Nicholas Michael, Peng Yang, Qingshan Liu, Dimitris Metaxas, and Carol Neidle. 2011. A framework for the recognition of nonmanual markers in segmented sequences of american sign language. In *Proceedings of the British Machine Vision Conference*, pages 124.1–124.12. BMVA Press.

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

OpenAI. 2024. Gpt-4o system card. ArXiv:2410.21276.

Pamela Perniss, Robin L. Thompson, and Gabriella Vigliocco. 2010. Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227.

Sina Proske, Annika Herrmann, Jana Hosemann, and Markus Steinbach. 2020. *A Grammar of German Sign Language (DGS)*, 1 edition. SIGN-HUB Sign Language Grammar Series. SIGN-HUB. Accessed 31-10-2021.

Markus Steinbach and Edgar Onea. 2015. A drt analysis of discourse referents and anaphora resolution in sign language. *Journal of Semantics*, 33(3):409–448.

William C Stokoe. 1980. Sign language structure. *Annual review of anthropology*, pages 365–390.

Fengfu Tsao. 1977. *A Functional Study of Topic in Chinese: The First Step Towards Discourse Analysis*. Ph.d. dissertation, University of Southern California.

Abhishek Bharadwaj Varanasi, Manjira Sinha, Tirthankar Dasgupta, and Charudatta Jadhav. 2024. Linguistically informed transformers for text to American Sign Language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 50–56, Bangkok, Thailand. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998. *Centering Theory in Discourse*. Oxford University Press.

Anne Wienholz, Derya Nuhbalaoglu, Nivedita Mani, Annika Herrmann, Edgar Onea, and Markus Steinbach. 2018. Pointing to the right side? An ERP study on anaphora resolution in German Sign Language. *PloS one*, 13(9):e0204223.

Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. 2021a. 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. Association for Computational Linguistics.

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

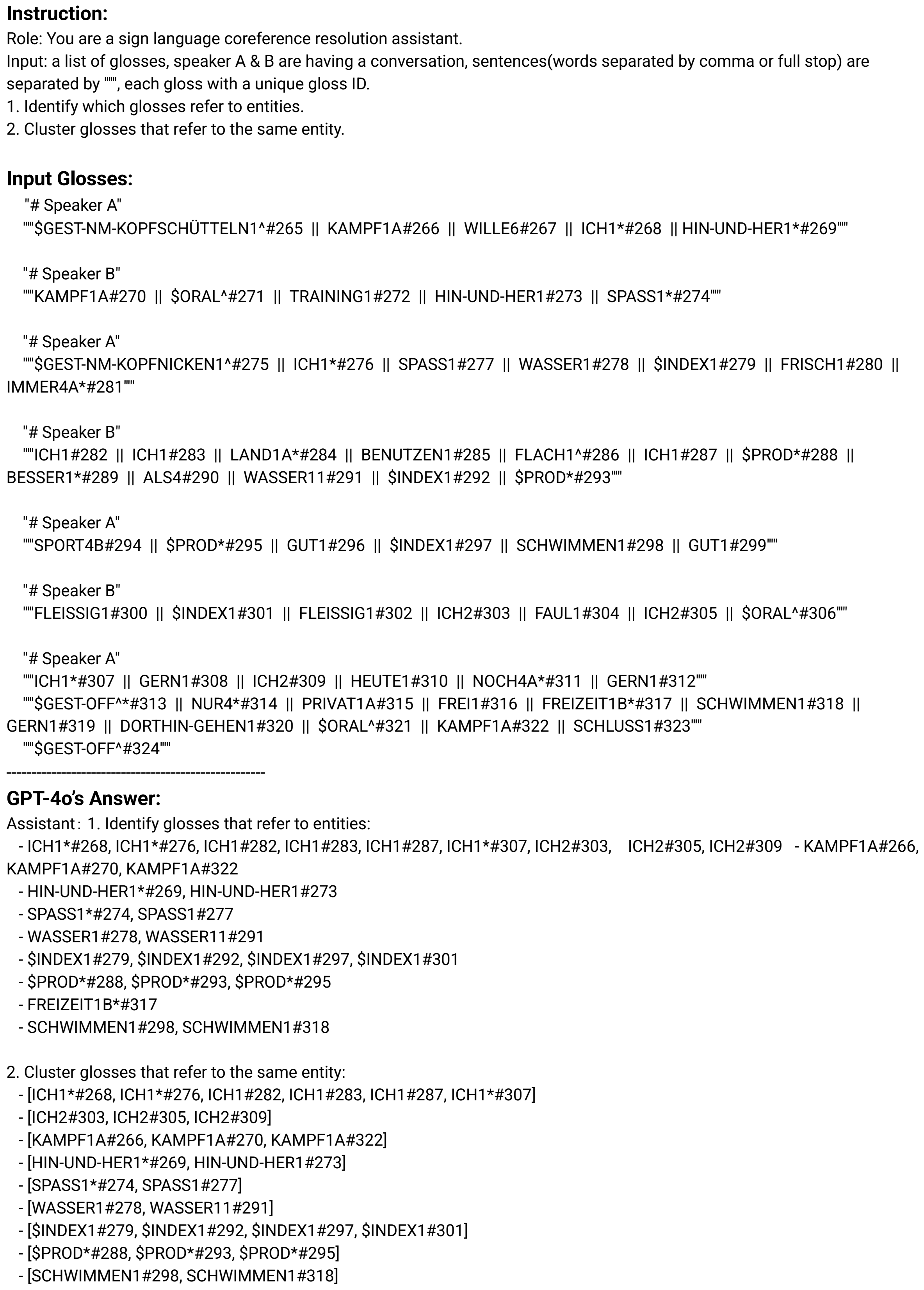
Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shulin Zhang, Jixing Li, and John Hale. 2022. Quantifying discourse support for omitted pronouns. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–12, Gyeongju, Republic of Korea. Association for Computational Linguistics.

# A Prompts

## A.1 An Example Prompt

Figure 3 shows a prompt we used in the experiment for *GPT-4o_zs*, along with the LLM's response. Note that the answer is summarized from the LLM's raw responses, which we manually checked.



**Instruction:**

Role: You are a sign language coreference resolution assistant.
Input: a list of glosses, speaker A & B are having a conversation, sentences(words separated by comma or full stop) are separated by """, each gloss with a unique gloss ID.
1. Identify which glosses refer to entities.
2. Cluster glosses that refer to the same entity.

**Input Glosses:**

"# Speaker A"
"""$GEST-NM-KOPFSCHÜTTELN1^#265 || KAMPF1A#266 || WILLE6#267 || ICH1*#268 || HIN-UND-HER1*#269"""

"# Speaker B"
"""KAMPF1A#270 || $ORAL^#271 || TRAINING1#272 || HIN-UND-HER1#273 || SPASS1*#274"""

"# Speaker A"
"""$GEST-NM-KOPFNICKEN1^#275 || ICH1*#276 || SPASS1#277 || WASSER1#278 || $INDEX1#279 || FRISCH1#280 || IMMER4A*#281"""

"# Speaker B"
"""ICH1#282 || ICH1#283 || LAND1A*#284 || BENUTZEN1#285 || FLACH1^#286 || ICH1#287 || $PROD*#288 || BESSER1*#289 || ALS4#290 || WASSER11#291 || $INDEX1#292 || $PROD*#293"""

"# Speaker A"
"""SPORT4B#294 || $PROD*#295 || GUT1#296 || $INDEX1#297 || SCHWIMMEN1#298 || GUT1#299"""

"# Speaker B"
"""FLEISSIG1#300 || $INDEX1#301 || FLEISSIG1#302 || ICH2#303 || FAUL1#304 || ICH2#305 || $ORAL^#306"""

"# Speaker A"
"""ICH1*#307 || GERN1#308 || ICH2#309 || HEUTE1#310 || NOCH4A*#311 || GERN1#312"""
"""$GEST-OFF^*#313 || NUR4*#314 || PRIVAT1A#315 || FREI1#316 || FREIZEIT1B*#317 || SCHWIMMEN1#318 || GERN1#319 || DORTHIN-GEHEN1#320 || $ORAL^#321 || KAMPF1A#322 || SCHLUSS1#323"""
"""$GEST-OFF^#324"""

---

**GPT-4o's Answer:**

Assistant: 1. Identify glosses that refer to entities:
- ICH1*#268, ICH1*#276, ICH1#282, ICH1#283, ICH1#287, ICH1*#307, ICH2#303, ICH2#305, ICH2#309 - KAMPF1A#266, KAMPF1A#270, KAMPF1A#322
- HIN-UND-HER1*#269, HIN-UND-HER1#273
- SPASS1*#274, SPASS1#277
- WASSER1#278, WASSER11#291
- $INDEX1#279, $INDEX1#292, $INDEX1#297, $INDEX1#301
- $PROD*#288, $PROD*#293, $PROD*#295
- FREIZEIT1B*#317
- SCHWIMMEN1#298, SCHWIMMEN1#318

2. Cluster glosses that refer to the same entity:
- [ICH1*#268, ICH1*#276, ICH1#282, ICH1#283, ICH1#287, ICH1*#307]
- [ICH2#303, ICH2#305, ICH2#309]
- [KAMPF1A#266, KAMPF1A#270, KAMPF1A#322]
- [HIN-UND-HER1*#269, HIN-UND-HER1#273]
- [SPASS1*#274, SPASS1#277]
- [WASSER1#278, WASSER11#291]
- [$INDEX1#279, $INDEX1#292, $INDEX1#297, $INDEX1#301]
- [$PROD*#288, $PROD*#293, $PROD*#295]
- [SCHWIMMEN1#298, SCHWIMMEN1#318]

Figure 3: An example prompt shows the instruction, input glosses, and the answer from *GPT-4o_zs*.