MMAR: Multilingual and multimodal anaphora resolution in instructional videos

Cennet Oguz, Pascal Denis, Simon Ostermann, Natalia Skachkova, Emmanuel Vincent, Josef van Genabith

▶ **To cite this version:**

Cennet Oguz, Pascal Denis, Simon Ostermann, Natalia Skachkova, Emmanuel Vincent, et al.. MMAR: Multilingual and multimodal anaphora resolution in instructional videos. Findings of the 2024 Conference on Empirical Methods in Natural Language Processing, Nov 2024, Miami, United States. hal-04733760

HAL Id: hal-04733760

https://inria.hal.science/hal-04733760v1

Submitted on 13 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

MMAR: Multilingual and Multimodal Anaphora Resolution in Instructional Videos

Cennet Oguz[1], **Pascal Denis**[2], **Simon Ostermann**[1], **Natalia Skachkova**[1]
Emmanuel Vincent[3] and **Josef van Genabith**[1]
[1]German Research Center for Artificial Intelligence (DFKI), Saarland Informatics
[2]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{cennet.oguz, ivana.kruijff, josef.van_genabith}@dfki.de
{pascal.denis, emmanuel.vincent}@inria.fr

Abstract

Multilingual anaphora resolution identifies referring expressions and implicit arguments in texts and links to antecedents that cover several languages. In the most challenging setting, cross-lingual anaphora resolution, training data, and test data are in different languages. As knowledge needs to be transferred across languages, this task is challenging, both in the multilingual and cross-lingual setting. We hypothesize that one way to alleviate some of the difficulty of the task is to include multimodal information in the form of images (i.e. frames extracted from instructional videos). Such visual inputs are by nature language agnostic, therefore cross- and multilingual anaphora resolution should benefit from visual information. In this paper, we provide the first multilingual and multimodal dataset annotated with anaphoric relations and present experimental results for end-to-end multimodal and multilingual anaphora resolution. Given gold mentions, multimodal features improve anaphora resolution results by $\sim 10\%$ for unseen languages.

1 Introduction

A procedural text is a sequence of instructions describing how to create or change an object in a certain way. Among the many genres of texts, procedural texts are the ones related to real-world applications such as robotics with human-robot interaction (Misra et al., 2016), video understanding of how-to videos (Miech et al., 2019; Zhukov et al., 2019; Mishra et al., 2021), etc. Procedural text understanding requires a system to track how a given entity undergoes change. However, using language for describing entity changes may raise linguistic ambiguities, which are a key challenge, especially in instructional videos. In particular, temporally evolving entities present rich and, to date, understudied challenges. Cooking recipes provide a paradigmatic source of potentially ambiguous referring expressions for ingredients that are undergoing physical or chemical changes such as chopping, boiling, mixing, etc. (Kiddon et al., 2015). Existing approaches include Huang et al. (2017, 2018), who suggest reference resolution, or Fang et al. (2022) and Oguz et al. (2022), who propose anaphora resolution for tracing the temporal change of entities in recipes.

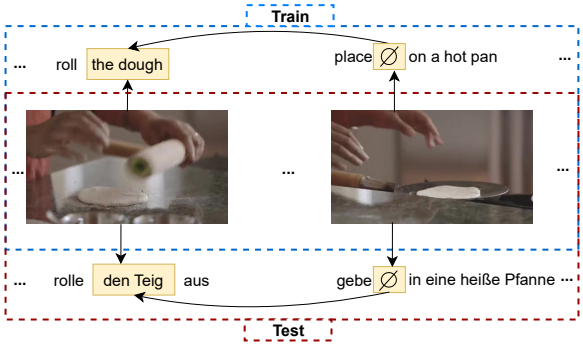

Figure 1: An example of multilingual and multimodal anaphora resolution data and learning strategy. The blue dashed box denotes the training phase, whereas the red dashed box shows the testing phase of our multilingual and multimodal anaphora resolution system. ϕ represents zero anaphors, also known as null or implicit arguments.

Anaphora resolution is the task of identifying the antecedent of an anaphor, where the antecedent is a language expression in the previous context that a given anaphor refers to (Poesio et al., 2018; Fang et al., 2022; Ye et al., 2023). Anaphoric language expressions can occur in many forms, such as sentences, nominal phrases including pronouns, and challenging zero anaphora (a.k.a, implicit arguments or null pronouns) which are not verbalized. An example of such a case is given in Figure 1, where the zero anaphora ϕ in the instruction *place* ϕ *on a hot pan* refers to the antecedent *the dough* in the previous context of the recipe. While anaphora resolution is classically seen as a uni-modal text-based task, there have been some attempts at adding modalities. Oguz et al. (2022) present the first attempt to formulate a multimodal anaphora reso-

	Multilingual	Multimodal	Parallel	Anaphorcity	Zero-Pronoun
PAWS (Nedoluzhko et al., 2018)	✓	✗	✓	C	✓
CorefUD 1.0 (Nedoluzhko et al., 2022)	✓	✗	✓	C	✓
ParCorFull2.0 (Lapshinova-Koltunski et al., 2022)	✓	✗	✓	C	✓
CRAC (Žabokrtský et al., 2022)	✓	✗	✗	C	✗
CIN (Goel et al., 2022)	✗	✓	✗	C	✗
Find-2-Find (Oguz et al., 2022)	✗	✓	✗	B,N,C	✓
VWP (Hong et al., 2023)	✗	✓	✗	C	✗
MMAR (ours)	✓	✓	✓	B,N,C	✓

Table 1: The list of the multilingual and multimodal anaphora resolution datasets shows the novelty of our collection. We use the ✓if the property is addressed by the corresponding data; if not, the mark is ✗. The *Anaphorcity* column indicates the anaphoric relations annotated in the data: B is Bridging, N is near-identity, and C is coreference.

lution task. Oguz et al. (2023a) present a system for anaphora resolution with object localization in a multitask learning setting to show the help of multimodal image information for anaphora resolution in recipe instructions of cooking videos. However, all previous work on multimodal anaphora resolution has only been done in English (Goel et al., 2022; Oguz et al., 2023b; Ates et al., 2023). Multilingual coreference resolution has been researched extensively: For example, Zhekova and Kübler (2010) show the importance of language-independent hand-crafted features, and Žabokrtskỳ et al. (2022, 2023) demonstrate the effectiveness of deep learning methods for multilingual coreference resolution. However, to the best of our knowledge, we have made the first attempt to investigate multimodal representations for multilingual anaphora resolution in procedural texts. We show that visual inputs provide important language-agnostic features and improve multilingual resolution.

Our contributions are as follows: (1) we provide novel multimodal parallel data[1] in English, Turkish, and German, which are annotated with anaphora resolution, including zero anaphora; (2) we provide a novel model architecture for multilingual and multimodal anaphora resolution with mention detection that outperforms a strong baseline (Oguz et al., 2022); (3) we show that multimodal representations outperform language-only multilingual representations on unseen data for anaphora resolution. To our knowledge, our dataset is the first and only multimodal parallel anaphora resolution dataset for multiple languages.

2 Related Work

Our work is at the intersection of two research areas: anaphora resolution and procedural text understanding. In this section, we summarize related work and emphasize the most important common points and differences between our data and already existing data sets, as well as our approach and previous methods.

2.1 Anaphora Resolution

Research on anaphora resolution is usually addressed in two different directions. The first one is the annotation task (Poesio and Artstein, 2008; Fang et al., 2022; Oguz et al., 2022; Ye et al., 2023) with a focus on determining mention types and anaphoric relations, i.e., coreference (Ghaddar and Langlais, 2016; Ng, 2017) and bridging (Rösiger, 2018; Poesio and Artstein, 2008), and identifying mentions (i.e., anaphor and antecedents) in a document that are involved in anaphoric relations. The second direction is the modeling of anaphora resolution (Lee et al., 2017, 2018; Joshi et al., 2020; Yu and Poesio, 2020; Pandit and Hou, 2021) to automatically identify anaphoric mentions and their anaphoric relations. Recent studies tackle anaphora resolution in an *end-to-end* fashion (Lee et al., 2017; Yu and Poesio, 2020) that jointly identifies the referring expressions and their anaphoric relations in a document. The state-of-the-art works focus on span-based language representation of mentions (Joshi et al., 2020) and Transformer-based resolution modeling (Pandit and Hou, 2021). Anaphora resolution has been studied separately from multilingual (Nedoluzhko et al., 2018; Žabokrtskỳ et al., 2022) and multimodal (Goel et al., 2022; Oguz et al., 2022) perspectives without connecting the two to date.

Multilingual Anaphora Resolution Multilingual anaphora resolution is an active area of research. The SemEval 2010 (Recasens et al., 2010) and CoNNL 2012 (Pradhan et al., 2012) shared tasks have contributed to progress in multilingual

[1] a placeholder for the link of code and data

anaphora resolution. ParCorFull2.0 (Lapshinova-Koltunski et al., 2022) is the only parallel multilingual coreference resolution data in German and English. Recently, the Universal Dependencies annotation scheme (De Marneffe et al., 2021) was used to synchronize datasets of different languages in the CorefUD 1.0 corpus (Nedoluzhko et al., 2022). This corpus was used for the CRAC shared tasks on multilingual coreference resolution (Žabokrtskỳ et al., 2022, 2023). The winning CorPipe system (Straka and Straková, 2022) implements a jointly trained pipeline approach based on the *end-to-end* system by Lee et al. (2017) to solve mention detection and then performs coreference linking on the retrieved mentions. To date, to the best of our knowledge, all existing multilingual datasets focus on coreference, and none considers anaphoric relations as bridging and near-identity, in contrast to our approach presented below.

Multimodal Anaphora Resolution Multimodal anaphora resolution combines language and visual representations for input to find the antecedent of the anaphor. Research on person or character identification in TV series or stories (Ramanathan et al., 2014; Cui et al., 2021; Hong et al., 2023; Liu and Keller, 2023) addressed the problem of coreference resolution in a given text. Similarly, Kong et al. (2014) and Goel et al. (2023) resolve coreferential referring expressions in more generic scenarios by using text and images as inputs. In contrast to previous studies, Oguz et al. (2022) focus on procedural texts from cooking videos that contain state changes of entities. The resolution complexity grows substantially with the increase in visual and textual complexity triggered by evolving entity changes. In summary, previous multilingual studies focus solely on coreference relations without multimodal information, while existing multimodal anaphora resolution studies investigate multimodal features only for English data. In the following, we propose a novel multilingual data set with visual inputs alongside the parallel multilingual cooking recipes for anaphora resolution (see Table 1).

3 Data

Language Data To create our data set, we started with available English multimodal anaphora resolution data and translated these data into more languages. In this study, we build on the Chop&Change anaphora resolution dataset and annotation schema of Oguz et al. (2022, 2023b) for two reasons. First, it includes near-identity anaphoric relations (unlike other approaches), bridging, and coreference for noun phrases, pronouns, and null pronouns (a.k.a, implicit arguments or zero anaphora). Second, cooking videos of the recipes are provided with annotations for video-instruction alignment. In total, in our work here we provide 400 multimodal training recipes and 100 multimodal test recipes in English; for data statistics, see Appendix A. We then manually translate each recipe into Turkish and German by native language speakers based on the original English recipe and video inputs, refer to A. As much as possible, we keep the sentence structure of the instructions while translating manually. For example, we do not drop the direct or prepositional objects or expand any zero anaphora in the source in the translations. Additionally, we use a pronoun in the translation whenever a pronoun is used for an entity. However, we also prepare and analyze pro-drop data in Turkish A. We keep the linguistic feature distribution as in the original English documents and manually apply the anaphora annotation to the other languages based on the annotation of English recipes. An example of a German translation and annotation of an English recipe is provided in Figure 2. In total, 1,200 training and 300 test recipes, a third each in English, German, and Turkish, are included in the multilingual and multimodal MMAR language dataset, with recipes in each language making up equal parts of the dataset. For detailed information regarding the translation and annotation of Turkish and German recipes, refer to Appendix A.

Visual Data For each video, procedural steps in a recipe are annotated with temporal boundaries and described by imperative instructions in the text. In other words, each recipe instruction is aligned to a part of the corresponding cooking video. For example, the instruction 'thinly slice the beef' in Figure 2 is annotated with the corresponding starting and ending times of the segment in the video. Following Oguz et al. (2022), each textual instruction is temporally aligned to a segment of the corresponding cooking video. Because of the mapping to a whole instruction video segment, a gold mention has no precise alignment to a specific frame. Thus, Zhou et al. (2018), Huang et al. (2018), Oguz et al. (2022), and Oguz et al. (2023a) randomly sample the frames for each instruction. In contrast in our work, we use the CLIP model (Radford et al., 2019) to select the best frame and the best region for the

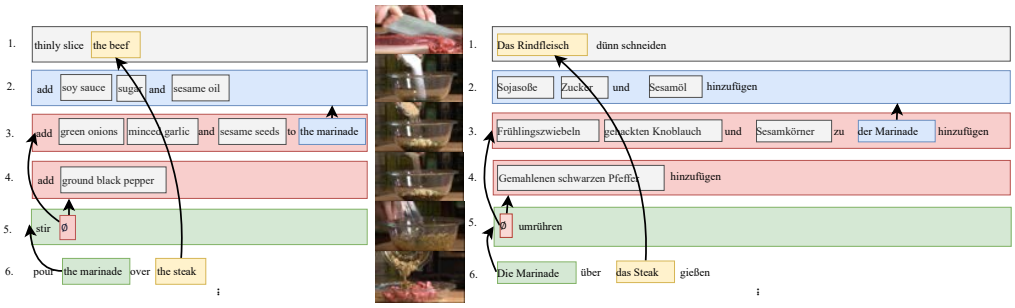

Figure 2: An example annotation of anaphora resolution for a Beef Bulgogi recipe in English and its translation in German with one frame of the video segments. The arrows start from the anaphor and point to the corresponding antecedent/s. Similarly, the anaphoric mentions are shown with the same color boxes. The gray boxes indicate singletons, i.e., mentions without any anaphoric links. We follow the same color coding as in English and translated German recipes. Turkish translation is in Figure 4.

instruction in a preprocessing step (i.e., CLIP is not trained with the rest of the model). To investigate the contribution of multimodal features, we train and evaluate our model with and without visual features of each gold and candidate mentions. We present the outcomes of two pre-trained video encoders used for extracting multimodal features in Appendix A.0.2.

4 Methodology

4.1 Task

The task of anaphora resolution is to assign each anaphoric mention (e.g., span) i in an instruction to one or more antecedents $y_i \in \{\epsilon, y_1, \ldots, y_{i-1}\}$, where ϵ is an (empty) dummy antecedent, and y_i is one of the preceding gold spans from the previous instructions. E.g., in Figure 2, the anaphor *the steak* in the 6th step refers to the antecedent *the beef* in the 1st step, and the null pronoun ϕ in the 5th step refers to the 3rd and the 4th steps. The selection of dummy ϵ as an antecedent indicates that the anaphor is either an incorrect span or a singleton without an antecedent such as *green onions* from the 3rd step in Figure 2.

4.2 Method

4.3 Baseline Model

We use an end-to-end anaphora resolution Chop&Change (C&C) system Oguz et al. (2022) based on end-to-end conference resolution Fang et al. (2021); Lee et al. (2017) as a baseline. This end-to-end resolution system begins with mention extraction, followed by anaphora resolution. The mention detection model is trained with the back-propagated anaphora resolution loss. There is no separate explicit loss that penalizes mention errors.

4.4 Proposed Model

We propose a multi-task learning method inspired by a hierarchically supervised multi-task learning model (Sanh et al., 2019) focused on semantic tasks (e.g., named entity recognition, relation extraction, coreference resolution, etc.) to learn enhanced word embeddings. A well-trained mention detection model is crucial for multi-lingual anaphora resolution. Our multitask learning method functions similarly to an end-to-end system, but it additionally backpropagates the mention loss to train the mention detection model together with the anaphora loss (in an alternating fashion, see Section 4.5 and Figure 3). Below, we provide a detailed formalization of our methods.

4.4.1 Input

Language Input. We consider all continuous token sequences with up to L words as a potential span. We use pre-trained XLM-RoBERTa (Conneau et al., 2019), a state-of-the-art multilingual model for a wide range of cross-lingual transfer tasks. Anaphoric language expressions can occur several sentences apart. Thus, capturing document-level dependencies is essential to judge whether two expressions are anaphoric (Pražák and Konopik, 2022). Let XLM-RoBERTa$(w_1, \ldots, w_T)$ be the word representations computed by the model, where w_1 is the first token and w_T is the last token of the overall recipe. A span x_i consists of one or more consecutive tokens of an instruction I_i in a recipe. We use the verb of an instruction as a pointer for null pronouns, e.g., *stir* is used as a pointer token for the null pronoun ϕ in the 5th instruction *stir* ϕ in Figure 2.

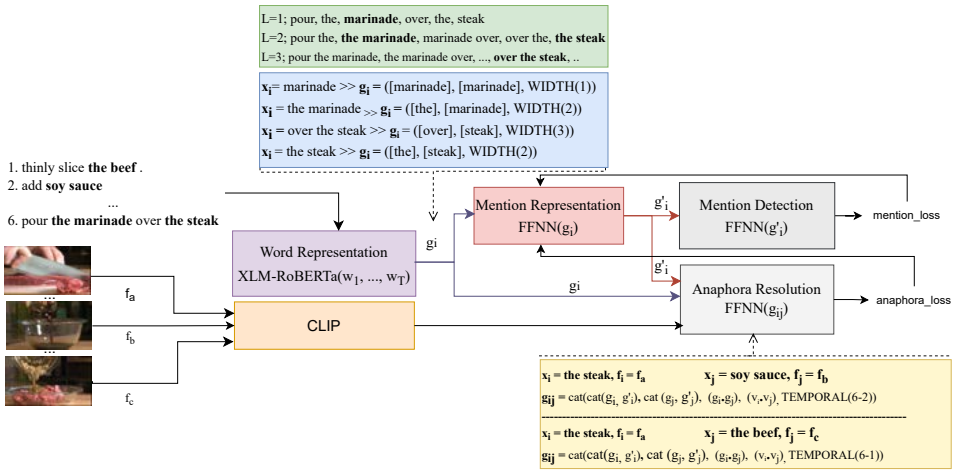

Figure 3: The model architecture for mention detection and anaphora resolution. md_loss is the mention detection loss, and ar_loss is the anaphora resolution loss. The green box shows the potential spans of token sequences with L tokens. The blue box shows some potential spans with representation for the mention detection module. The yellow box shows the pair representation for anaphora resolution. f_a, f_b, and f_c represent the frames of the spans *the beef*, *soy sauce*, and *the steak*, respectively. A word within a bracket denotes the vector of the words, e.g., [the] is the vector of *the*. cat(.) means the concatenation of the given inputs.

Visual Input. Each cooking video consists of more than one segment, each corresponding to one instruction. Each segment consists of many frames. We pick one frame to represent each instruction. Each frame f is encoded using the CLIP model (Radford et al., 2021). The frame-level vectors obtain the instruction's visual feature vector: $v_i = \text{CLIP}(f)$ for the spans extracted from the instruction I_i, as Figure 3 illustrates.

4.4.2 Mention Detection

For mention detection, following Lee et al. (2017) and Oguz et al. (2022), we consider potential spans and compute the corresponding span score. Up to L, we consider all the continuous sequences of tokens as potential spans for instructions (see green box in Figure 3). XLM-RoBERTa is used to extract the contextualized multi-lingual word embeddings $x_t^* = \text{XLM-RoBERTa}(w_1, \ldots, w_T)$ where x_t^* refers to the vector representation of the token at time t of the recipe. The vector representation g_i of a given span is obtained by concatenating the word vectors of its boundary tokens and its width:

$$g_i = [x^*_{\text{START}(i)}, x^*_{\text{END}(i)}, \phi(i)], g'_i = \text{FFNN}(g_i) \quad (1)$$

$\text{START}(i)$ and $\text{END}(i)$ represent the starting and ending token indexes for g_i, respectively. $\phi(i) = \text{WIDTH}(\text{END}(i) - \text{START}(i))$ is the width feature of the span where $\text{WIDTH}(.)$ is the embedding function of the predefined bins of $[1, 2, 3, 4, 8, 16]$. An example is provided in the blue box in Figure 3, where x_i = *over the steak* is represented with the concatenation of the vector of start token *over*, the end token *steak*, and $\text{WIDTH}(3)$. The mention score is computed for each span using cross-entropy loss, and the mention model is trained with back-propagated mention loss mention_loss = $\textit{softmax}(\text{FFNN}(g'_i))$ as in Figure 3.

4.4.3 Anaphora Resolution

Following Oguz et al. (2022), we formulate anaphora resolution as a classification task over pairs of gold and candidate mentions. The representation of a span pair g_{ij} is obtained by concatenating the contextualized representations (i.e., equation 1) two span embeddings g_i and g_j, and element-wise multiplication of the corresponding span representation, $g_i \cdot g_j$:

$$g_{ij} = [g_i, g_j, g_i \cdot g_j, \phi_{temp}(i, j))].$$

$$\phi_{temp}(i, j) = \text{TEMPORAL}(\#a_j - \#a_i)$$

where the feature vector $\phi_{temp}(i, j)$ is the distance between the step index of span i and span j, $\text{TEMPORAL}(\cdot)$ (Oguz et al., 2022) is an embedding function that uses the predefined list of bins of $[1, 2, 3.., 30]$, $\#a_i$ refers to the instruction index of span i and $\#a_j$ to the instruction index of span j. We concatenate $\phi_{temp}(i, j)$ to obtain the vector representation of a span pair. We minimize the cross entropy loss for candidate span pairs with anaphora_loss = $\textit{sigmoid}(\text{FFNN}(g_{ij}))$ for resolution, as in Figure 3.

Visual Features in Anaphora Resolution Oguz et al. (2022) assumes two spans are likely to be anaphoric if their frames are similar and define the visual similarity of pairs via the dot product of the paired visual features, $v_i \cdot v_j$, in the anaphora resolution (cf. Figure 3), referred to Frame-Cos:

$$g_{ij} = [g_i, g_j, g_i \cdot g_j, v_i \cdot v_j, \phi_{temp}(i,j))]. \quad (2)$$

We instead instead concatenate the visual features with the span language features $[g_i, v_i]$ and $[g_j, v_j]$ in the representation of g_{ij} to condition the span representation to language-agnostic visual features, referred to Frame-Span:

$$g_{ij} = [g_i, v_i, g_j, v_j, g_i \cdot g_j, \phi_{temp}(i,j))]. \quad (3)$$

4.5 Training Details

We chose the straightforward yet highly effective training detailed in (Søgaard and Goldberg, 2016; Sanh et al., 2019): at each epoch in the training process, a task (i.e., mention detection or anaphora resolution) is chosen randomly, and we select a batch of training data for the assigned task to update parameters. This cycle continues until convergence is achieved. Hence, the mention representation layer learns using mention and anaphora errors.

4.6 Evaluation

Following Hou et al. (2018) and Oguz et al. (2022), we analyze the performance of our end-to-end hierarchical anaphora resolution model with the F1-score, where precision is the result of dividing the number of correctly predicted pairs by the total number of predicted pairs and recall is computed by dividing the number of correctly predicted pairs by the total number of gold pairs.

5 Experiments

We conduct various cross- and multi-lingual experiments to investigate the impact of visual features on anaphora resolution. The main body of the paper focuses solely on English monolingual experiments incorporating multi-lingual and frame features, while the Appendix B details experiments involving Turkish and German, along with video encoders.

5.1 Gold and Candidate Mentions

We define candidate mentions to be all possible continuous token sequences (Clark and Manning, 2016; Lee et al., 2017). In Figure 2, for example, *to the marinade* or *seeds to the marinade* are examples of candidate mentions in the 3*rd* step, that, however, do not correspond to correct spans. In contrast, gold mentions are all correct spans, e.g., *sesame seeds* and *the marinades* in the 3*rd* step in Figure 2. When gold mentions are given, a mention detection module is unnecessary and thus left out of the gold mention-based experiments, whereas candidate mentions heavily depend on mention detection. Therefore, we conduct separate experiments for the candidate and gold mentions to investigate the effect of multimodal features.

Gold Mentions. We have gold mentions for each language in both train and test datasets. Thus, we train our model without mention detection component with/out multimodal features for cross- and multi-lingual experiments. Note that mention detection is unnecessary for anaphora resolution when using gold mentions. Our model, without vision and using Frame-Cos, is equivalent to the C&C model (Oguz et al., 2022) and serves as baseline.

Candidate Mentions. To apply anaphora resolution for candidate mentions, we first need to detect the mentions (e.g., extracting correct spans from the recipe instructions). Thus, we add a mention representation and detection layers for capturing the mention features trained for the mention prediction $[g'_i, g'_j]$ to Equation 2 and Equation 3 as in the yellow box in Figure 3. Additionally, our multitask training approach allows us to train mention detection separately from anaphora resolution. We conducted experiments with multi-lingual mention detection, training the mention detection component using English, Turkish, and German mentions, while training the anaphora resolution in English.

5.2 Multi-lingual and Multimodal Features

To check the benefit of cross- and multi-lingual features for anaphora resolution, we compare the results of multi-lingual experiments with cross-lingual experiments. Multimodal features might provide language-agnostic knowledge of anaphora resolution for the challenging task of multi-lingual anaphora resolution for unseen languages. However, the best way to use visual features for anaphora resolution is an open question. To investigate the visual features for both anaphora resolution and mention representation, we focus on two methods: (1) *Frame-Cos* (Oguz et al., 2022), where we use the similarity of visual representations of instructions for mention pairs in the anaphora resolu-

		Turkish			German			English		
		NPs	Zero	Full	NPs	Zero	Full	NPs	Zero	Full
Trained on English	no vision	25.41	12.42	22.22	38.10	22.91	33.56	50.35	67.84	56.19
	Frame-Cos	26.86	33.73	28.93	37.80	39.54	38.43	51.66	70.62	57.88
	Frame-Span	**31.95**	**43.94**	**35.63**	**41.16**	**49.45**	**44.18**	**53.52**	**71.43**	**59.34**
Trained on Multi-lingual	no vision	44.86	54.57	47.89	45.39	61.33	52.13	50.47	67.97	56.41
	Frame-Cos	50.47	61.37	53.76	49.11	70.80	58.18	52.66	69.97	58.16
	Frame-Span	**50.57**	**64.36**	**54.72**	**51.82**	**72.16**	**60.34**	**54.39**	**72.36**	**60.21**

Table 2: F1 scores of the anaphora resolution for gold mentions, for models without vision features and with both *Frame-Cos* and *Frame-Span*. The model is trained with only English (English Spans) and multi-lingual data with English, German, and Turkish (Multi-lingual Spans) as well as with multimodal features in Frame-Cos and Frame-Span methods.

tion module (see Equation 2), and (2) *Frame-Span*, where we learn mention representations using the corresponding instruction visual features directly (see Equation 3).

6 Results and Discussion

In Table 2 3, and 4, we compare the multi-lingual experiments with the English-only results. Appendix Section B and Table 7 provide more detailed results and experiments, including both language-visual perspectives and mention detection.

Overview. We present the anaphora resolution results in the cooking domain with multimodal and multi-lingual features for both candidate and gold mentions in English, German, and Turkish, in two settings: Training on mono-lingual data only, i.e., a *cross-lingual* setting, and training on all three languages simultaneously, i.e., a *multi-lingual* setting. Overall, the results with candidate mentions in Table 3 / Table 4 and gold mentions in Table 2 show that visual features improve both nominal and zero anaphora resolution results for seen and unseen languages, especially for zero anaphora resolution, both for cross- and multi-lingual settings. German and English are from the same language family, whereas Turkish is from a different family (the Altaic family). Therefore, for both candidate and gold mentions, we observe a similar pattern in anaphora resolution results according to language family relatedness: For a model trained on English data only, the transfer to German works better than to Turkish, exemplified by the better cross-lingual results for German. Furthermore, the performance for candidate mentions is propagated to subsequent tasks due to the sequential structure of the hierarchical system, as shown in Figure 3: The difference between the results given candidate and gold spans demonstrates that the mention detection model for the candidate mentions propagates errors to the actual anaphora resolution. As an example, if the noun phrase *the marinade* in the 3*th* step in Figure 2 is not detected as a mention by the mention detection model, the anaphora resolution model fails to detect *the marinade* as mention to resolve the antecedent. In Table 2, our Frame-Cos and no-vision models are equivalent to the baseline, i.e., the C&C model. We present further results utilizing models trained on Turkish and German data. Additionally, we also employ video features to investigate the impact of video encoders on multi-lingual anaphora resolution, refer to Appendix B.

Multi-lingual Features. We evaluate the effectiveness of multi-lingual features using two approaches. The first approach involves multi-lingual mention detection combined with English anaphora resolution (Table 3). Our proposed model can learn multi-lingual mention features through distinct mention detection processes that do not require anaphora annotation, refer to 7. This scenario is particularly prevalent, as it is common to have data available for mention detection (e.g., span detection) in various languages without corresponding anaphora annotations. The second approach includes both multi-lingual mention detection and multi-lingual anaphora resolution (Table 4). Table 3 shows that multi-lingual features in mention detection are effective for anaphora resolution in unseen languages but inefficient for anaphora resolution in the seen English language. Table 4 shows that multi-lingual training considerably improves both nominal and zero anaphora resolution results for Turkish and German, given candidate mentions. In contrast, cross-lingual cases with training in English are much more challenging. A slight improvement is only achieved in the setting with visual features for English. Similar tendencies can be observed for the anaphora resolution given gold mentions in Table 2 and 7. The slight improve-

		Turkish			German			English		
		NPs	Zero	Full	NPs	Zero	Full	NPs	Zero	Full
C&C English	Frame-Cos	4.11	5.88	4.64	4.59	0.41	2.99	39.29	57.18	44.25
Trained	no vision	4.26	2.23	3.68	8.81	2.77	6.65	40.07	58.58	45.44
on	Frame-Cos	6.45	5.32	6.11	8.40	5.53	7.28	43.33	57.61	47.88
English	Frame-Span	**7.30**	**16.98**	**10.37**	**9.17**	**15.87**	**11.86**	**44.02**	**64.43**	**50.33**
Multi-lingual Mention	Frame-Cos	13.99	6.10	11.71	27.26	13.58	22.31	42.06	58.33	47.12
English Anaphora	Frame-Span	**19.12**	**24.44**	**20.72**	**28.60**	**28.80**	**28.67**	**42.54**	**61.71**	**48.57**

Table 3: F1 scores for candidate mentions for the anaphora resolution of baseline C&C English (Oguz et al., 2022) and our models with/out vision features with *Frame-Cos* and *Frame-Span*. The model is trained only in English and has multimodal features in Frame-Cos and Frame-Span methods. *Multi-lingual Mention / English Anaphora* is trained on multi-lingual mention detection, while the anaphora resolution component is trained exclusively in English. The models are trained for 500 epochs.

		Turkish			German			English		
		NPs	Zero	Full	NPs	Zero	Full	NPs	Zero	Full
C&C Multi-lingual	Frame-Cos	33.66	47.76	37.98	40.53	53.33	45.85	43.07	53.15	46.24
Trained	no vision	32.18	49.45	37.76	39.66	54.67	45.83	42.76	58.98	47.96
on	Frame-Cos	34.85	48.13	39.02	40.19	55.64	46.62	42.92	60.67	48.87
Multi-lingual	Frame-Span	**37.29**	**50.96**	**41.51**	**40.87**	**55.95**	**47.05**	**45.50**	**61.05**	**50.56**

Table 4: F1 scores of the anaphora resolution for candidate mentions of baseline C&C multi-lingual and our models with/out vision features and with both *Frame-Cos* and *Frame-Span*. The model is trained with multi-lingual data in English, German, and Turkish as well as with multimodal features in Frame-Cos and Frame-Span methods. The models are trained for 500 epochs.

ment with multi-lingual features of gold mentions of English comes with zero anaphora resolution, whereas there is no apparent effect on the nominal anaphora resolution.

Multimodal Features. According to Tables 2, 3 and 4, multimodal features improve the results of anaphora resolution for German, Turkish, and English for both seen and unseen languages. The influence of visual features on anaphora resolution results of gold mentions, in Table 2, is clearly apparent with English training. When we test the model trained with English, we observe a minimum 4% improvement for nominal anaphora resolution and a minimum 10.4% improvement for zero anaphora resolution for unseen Turkish and German gold mentions. English results, on the other hand, show a slight improvement with overall anaphora resolution. When investigating the correct predictions, we find that the multimodal features are highly effective, especially for the resolution of zero and pronominal anaphors for seen and unseen languages. However, we find indications that the effect of the visual features decreases when the distance between the anaphor and antecedent pair is increased. For the German and English examples in Figure 2, visual features are helpful to predict the 4*th* instruction as the antecedent of zero anaphora ϕ in the 5*th*; however, the model fails to predict the 3*th* instruction as antecedent. On the other hand, the impact of the visual features is obscure for unseen languages for candidate mentions (s. Table 4), whereas w We observe a slight improvement with the Frame-Span method for multilingual Turkish, German, and English. We conjecture that visual features induce unwarranted similarity between correct and incorrect spans of mentions. For example, we obtain the same visual input for incorrect span *marinade over* and correct span *the marinade*.

7 Conclusion and Future Work

In this study, we presented a novel multilingual and multimodal dataset for anaphora resolution. Additionally, we empirically validated the benefit of multimodal features on anaphora resolution in a multilingual setting. We hope that our work will help to show that using additional language-agnostic modalities, such as the vision modality, can help to bridge gaps between languages, especially if training data are sparse or unavailable. Our results also indicate that mention detection is a bottleneck for anaphora resolution for unseen languages that conceals the positive impact of visual features. Thus, span-based language models like SpanBert (Joshi et al., 2020) need to be developed for multi-lingual settings.

Limitations

The primary limitation of our studies is the pre-trained visual and language model employed. Utilizing a more robust multilingual language model could further enhance the results. However, larger models may not always be feasible, and our method offers new avenues for improvement when model size is constrained. Furthermore, our focus on the cooking domain provides a controlled environment for the anaphora resolution task, but the applicability of our results to broader domains remains to be validated.

Ethics Statement

The videos of our data are publicly available and have already been published. Thus, our study does not create risk or privacy problems.

References

Halim Cagri Ates, Shruti Bhargava, Site Li, Jiarui Lu, Siddhardha Maddula, Joel Ruben Antony Moniz, Anil Kumar Nalamalapu, Roman Hoang Nguyen, Melis Ozyildirim, Alkesh Patel, et al. 2023. Marrs: Multimodal reference resolution system. *arXiv preprint arXiv:2311.01650*.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. 2021. Who's waldo? Linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Elsa Eiriksdottir and Richard Catrambone. 2011. Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human factors*, 53(6):749–770.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1362–1375, Online. Association for Computational Linguistics.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2022. Who are you referring to? Weakly supervised coreference resolution with multimodal grounding. *arXiv preprint arXiv:2211.14563*.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023. Semi-supervised multimodal coreference resolution in image narrations. *arXiv preprint arXiv:2310.13619*.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "it": Weakly-supervised, reference-aware visual grounding in instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2183–2192.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992.

Laura Ellen Knecht. 1985. *Subject and object in Turkish*. Ph.D. thesis, Massachusetts Institute of Technology.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? Text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. Parcorfull 2.0: A parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 805–813.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Danyang Liu and Frank Keller. 2023. Detecting and grounding important characters in visual stories.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Dipendra K. Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2016. Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300.

Anna Nedoluzhko, Michal Novák, and Maciej Ogrodniczuk. 2018. PAWS: A multi-lingual parallel treebank with anaphoric relations. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 68–76, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Vincent Ng. 2017. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Proceedings of the aaai conference on artificial intelligence*, volume 31.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer.

Cennet Oguz, Pascal Denis, Emmanuel Vincent, Simon Ostermann, and Josef van Genabith. 2023a. Find-2-find: Multitask learning for anaphora resolution and object localization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8099–8110, Singapore. Association for Computational Linguistics.

Cennet Oguz, Pascal Denis, Emmanuel Vincent, Simon Ostermann, and Josef van Genabith. 2023b. Find-2-find: Multitask learning for anaphora resolution and object localization. In *2023 Conference on Empirical Methods in Natural Language Processing*.

Cennet Oguz, Ivana Kruijff-Korbayová, Pascal Denis, Emmanuel Vincent, and Josef van Genabith. 2022. Chop and change: Anaphora resolution in instructional cooking videos.

Elif Oral, Ali Acar, and Gülşen Eryiğit. 2024. Abstract meaning representation of turkish. *Natural Language Engineering*, 30(1):171–200.

Onkar Pandit and Yufang Hou. 2021. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163, Online. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New

Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40.

Ondřej Pražák and Miloslav Konopik. 2022. End-to-end multilingual coreference resolution with mention head prediction. *arXiv preprint arXiv:2209.12516*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking people in videos with "their" names using coreference resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 95–110. Springer.

Marta Recasens, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 1–8.

Ina Rösiger. 2018. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6949–6956.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Atsushi Ushiku, Hayato Hashimoto, Atsushi Hashimoto, and Shinsuke Mori. 2017. Procedural text generation from an execution video. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 326–335, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Klaus Von Heusinger and Jaklin Kornfilt. 2005. The case of the direct object in turkish: Semantics, syntax and morphology. *Turkic languages*, 9(3):44.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Frank F Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. 2020. A benchmark for structured procedural knowledge extraction from cooking videos. *arXiv preprint arXiv:2005.00706*.

Bingyang Ye, Jingxuan Tu, and James Pustejovsky. 2023. Scalar anaphora: Annotating degrees of coreference in text. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 28–38.

Juntao Yu and Massimo Poesio. 2020. Multi-task learning based neural bridging reference resolution. *arXiv preprint arXiv:2003.03666*.

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards universal segmentations: UniSegments 1.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1137–1149, Marseille, France. European Language Resources Association.

Zdeněk Žabokrtskỳ, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 1–15.

Zdeněk Žabokrtskỳ, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. *arXiv preprint arXiv:2209.07841*.

Desislava Zhekova and Sandra Kübler. 2010. UBIU: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99.

Luowei Zhou, Nathan Louis, and Jason J. Corso. 2018. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Data

Procedural texts consist of instructions that direct individuals on how to carry out procedural tasks. They explain the necessary steps or rules to follow to complete the task (Eiriksdottir and Catrambone, 2011; Ushiku et al., 2017; Xu et al., 2020). Anaphora resolution for procedural text is challenging because of a severe lack of annotated data. The sheer amount of diverse actions and temporally evolving entities amplifies the difficulty of annotating sequences of instructions with anaphoric relations. Only a few datasets have been released for anaphora resolution in procedural texts, including a corpus by Fang et al. (2021) belonging to the chemistry domain and datasets from the cooking domain by Fang et al. (2022) and Oguz et al. (2022, 2023b). However, there is currently a lack of multilingual procedural texts that have been annotated with anaphora resolution. We consider the Chop&Change dataset Oguz et al. (2022, 2023b) which includes video and text of cooking recipes; refer Table 5 for the statistics of text recipes of Chop&Change. Hence, we manually translate the Chop&Change dataset in German and Turkish by native language annotators based on the original English recipe and video inputs. We select Turkish to analyze the impact of multimodal features on multilingual anaphora resolution when the unseen language belongs to a different language family. Conversely, we choose German to examine the effect of multimodal features on multilingual anaphora resolution when the languages belongs to the same language family. Then, we manually extend the anaphoric annotation to encompass German and Turkish recipes; see Figure 2 and Figure 4. In this section, we elucidate the process of translating and annotating German and Turkish recipes.

A.0.1 Turkish Dataset

When comparing Turkish to English and German, we observe that Turkish is a highly agglutinative language. Thus, suffixes play a critical role in Turkish grammar. To form a cooking instruction, the appropriate suffix must be added to the verb root based on the subject of the command. Thus, this section explains our English-Turkish translation process, focusing on imperative verbs and accusative nouns.

	Train	Test
Entities	9,213	2, 893
Zero Anaphor	997	266
Pronoun	304	139
Nominal	7,912	2,488
Pairs	4,715	1,485
Zero Anaphor-Antecedent	1,621	449
Pronoun Anaphor-Antecedent	368	165
Nominal Anaphor-Antecedent	2,726	871
Instruction	4,582	1,436
Recipe	400	100

Table 5: Annotation statistics of MMAR English data.

Nouns. Turkish, like many languages, utilizes grammatical cases to convey the function of nouns and pronouns within sentences. These cases play a crucial role in indicating various relationships and contexts(Knecht, 1985). The accusative case marks the verb's direct object and is formed by adding specific suffixes to the noun. For example, "the beef" in instruction 1 of our example in Figure 4 is translated as "dana etini." To make "dana eti" (beef) accusative, we add the suffix "-ni," resulting in "beef-ACC." Meanwhile, the dative case signifies the indirect object and often translates to "to" or "for" in English (Knecht, 1985; Von Heusinger and Kornfilt, 2005) and formed by adding the suffix "-e, -a" to the noun. For example, "to the marinade" (step 3 in Figure 4)is translated into Turkish as "marineye." Here, we append the suffix "-ye" to "marine" to form "marinade-DAT" in accordance with the dative case. Other cases include the locative and instrumental for serving distinct functions such as indicating location, possession, or means of action.

Pronouns. Turkish is a pro-drop language, which typically omits subject and object pronouns (Knecht, 1985; Oral et al., 2024). Turkish omit object pronouns in sentences where the context makes the object transparent. Therefore, we prepared two sets of Turkish-translated data: one where pronouns are omitted and another where the pronouns are retained as they are in the original English text. In Figure 5(b), We omit the pronoun "onu" (red-colored token, it-ACC) in the Turkish translation, replacing it with a zero-pronoun (text in green-box)

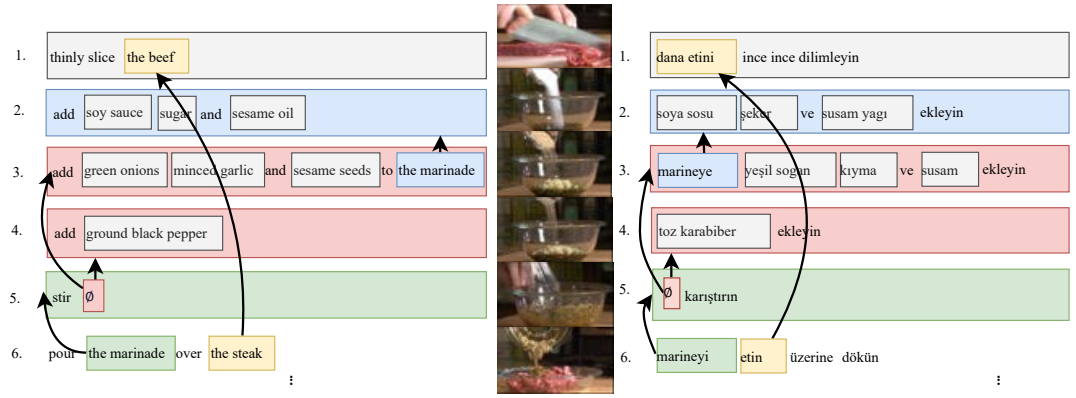

Figure 4: An example annotation of anaphora resolution for a Beef Bulgogi recipe in English and Turkish with one frame of the video segments. The arrows start from the anaphor and point to the corresponding antecedent/s. Similarly, the anaphoric mentions are shown with the same color boxes. The gray boxes indicate singletons, i.e., mentions without any antecedents. We follow the same color coding as in English and translated Turkish recipes.

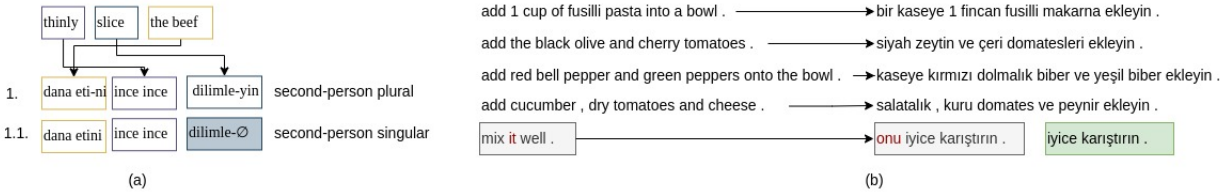

Figure 5: Turkish translation for verb and pronouns: (a) illustrates how verbs are used to construct polite imperative sentences in Turkish (b) demonstrates the phenomenon of pronoun omitting in Turkish translation.

because the context clearly expresses the object itself. Then, we investigate the effect of multimodal features on both sets of Turkish data; see the results of Turkish with omitted pronouns in Table 8.

Verbs. In the imperative form of Turkish, the second-person singular uses the bare verb stem without the infinitive ending. However, other imperative forms use various suffixes. Additionally, the imperative mood receives suffixes that vary depending on the formality of the command and the subject pronoun. For example, the verbal stem "dilimle" in the 1.2 example of Figure 5 is the second-person singular informal imperatives that are frequently employed to issue commands or make requests. Therefore, we selected the second-person plural to convey the courteous and instructive ambiance typical of cooking videos in our annotation; see the first example "dilimle-yin," in Figure 5. Additionally, the choice of vowel in suffixes depends on vowel harmony rules and the final vowel of the verb root. For example, "dök-ün" in sixth step and "dilimley-in" in the first step.

A.0.2 Visual Dataset

Various approaches can be used for video vector representation, such as frame sampling (Zhou et al., 2018; Huang et al., 2018), spatio-temporal modeling (Ni et al., 2022; Wang et al., 2022), sequential processing, and hybrid methods. The choice of approach depends on the specific application and the computational resources available. We use three different approaches: (1) Video-Git[2] (Wang et al., 2022), which is video encoding based on a generative (e.g., caption generation) decoding (2) Video-XCLIP[3] (Ni et al., 2022) for the discriminative method (Ni et al., 2022), which is video encoding based on a discriminative video classification method, (3) frame sampling based on the best frame selection with CLIP (Radford et al., 2021), see in Figure 6. Each cooking video is divided into multiple segments, each representing a single instruction. Each segment is made up of numerous frames, see in Figure 6. Thus, we have two potential approaches: using video segments con-

[2] https://huggingface.co/docs/transformers/main/en/model_doc/git

[3] https://huggingface.co/docs/transformers/main/en/model_doc/xclip

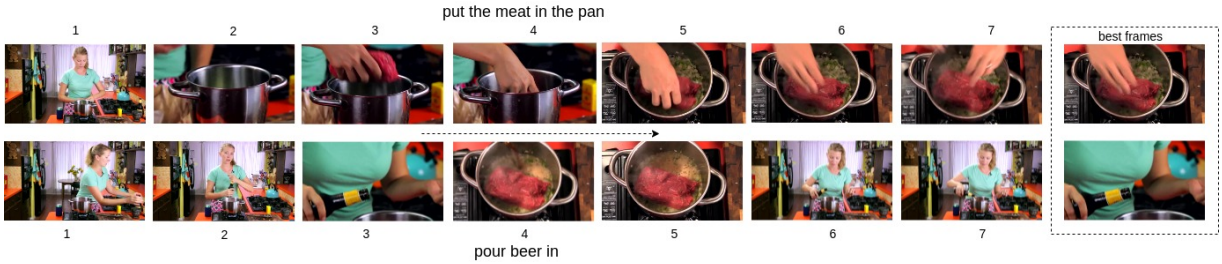

Figure 6: Frames of video segments of two instructions from a recipe. The frame numbers depict the order of frames. The best frames are selected based on the similarity score from the CLIP model (Radford et al., 2021), explained in Section 3.

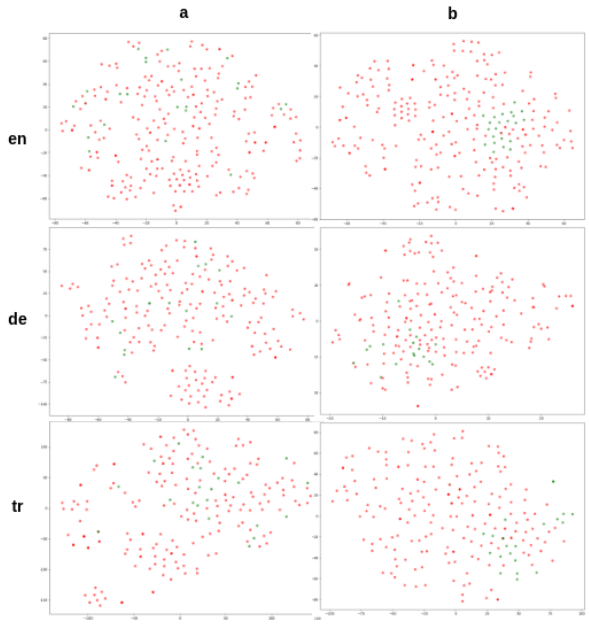

Figure 7: Visualization of mention representation in a recipe for English (en), German (de), and Turkish (tr) is shown with gold mentions in green and incorrect mentions in red, both before (a) and after (b) mention detection training on a multilingual anaphora resolution dataset.

taining multiple frames or selecting the best single frame to represent the visual features. In this study, we utilize video and image (for our frames) encoders. For video encoding, we provide the video segment corresponding to the instruction to the video encoder (Video-Git and Video-XCLIP) and apply mean-pooling to represent the video as a one-dimensional vector. To extract the best frame, we pick the best frame of a video that best matches a given description using CLIP (Radford et al., 2021) model, see Figure 6. CLIP can assess the similarity between images and text descriptions, enabling us to score each frame against the provided description and choose the best match. Therefore, we encode all video segment frames and the instruction using CLIP, compute the cosine similarity between each frame's embedding and the text embedding, and select the frame with the highest similarity score.

	Turkish	German	English
Trained on Turkish	78.14	52.34	56.62
Trained on German	39.78	78.15	29.36
Trained on English	45.58	45.70	78.86
Trained on Multilingual	**79.61**	**79.08**	**82.02**

Table 6: F1 scores of mention detection results for candidate mentions' cross- and multilingual features.

B Additional Results

We are not able to present all results in Table 3, 4, and 2 because of page limitations. Thus, we use this to expand our results of multimodal and multilingual experiments to show the effect of our model and multimodal features on multilingual anaphora resolution of the cooking domain.

B.1 Mention Detection

Mention detection is a crucial element in anaphora resolution systems, tasked with identifying mentions, including nominals, pronominals, and zero anaphora. We test the mention detection component of our model trained with cross- and multilingual datasets. To evaluate, we apply the F1 score, where precision is the result of dividing the number of correctly predicted mentions by the total number of predicted mentions, and recall is computed by dividing the number of correctly predicted mentions by the total number of gold mentions. Table 6 shows the cross- and multi-lingual mention detection results. Despite German and English belonging to the same language family, the model trained on Turkish data outperforms the model trained on German data for recognizing unseen English men-

		Turkish			German			English		
		NPs	Zero	Full	NPs	Zero	Full	NPs	Zero	Full
Trained on Turkish	no vision	46.78	58.06	50.23	27.99	53.32	35.75	35.03	61.28	42.22
	Video-GIT	47.33	59.06	50.93	28.05	54.73	36.69	35.54	61.5	43.59
	Video-XCLIP	47.1	59.27	50.83	29.31	53.83	37.31	36.92	62.91	44.94
	Frame-Cos	48.30	62.16	52.06	34.75	60.02	43.85	39.87	66.83	48.20
	Frame-Span	**49.71**	**64.13**	**54.12**	**34.86**	**64.85**	**46.72**	**39.92**	**68.77**	**48.49**
Trained on German	no vision	32.60	49.86	37.11	47.08	64.52	54.45	38.07	62.64	45.03
	Video-GIT	33.50	50.82	37.94	47.02	65.30	54.77	39.35	63.18	45.55
	Video-XCLIP	33.28	50.72	37.90	46.90	65.23	54.50	38.71	62.53	45.06
	Frame-Cos	35.62	54.02	40.34	48.75	66.95	56.15	39.76	64.25	46.58
	Frame-Span	**37.77**	**55.05**	**42.35**	**50.74**	**69.06**	**58.53**	**42.07**	**66.34**	**48.96**
Trained on English	no vision	25.41	12.42	22.22	38.10	22.91	33.56	50.35	67.84	56.19
	Video-GIT	25.21	23.12	24.64	37.55	24.03	33.11	51.75	68.81	57.29
	Video-XCLIP	26.73	20.08	24.95	36.01	34.67	35.56	51.49	70.71	57.92
	Frame-Cos	26.86	33.73	28.93	37.80	39.54	38.43	51.66	70.62	57.88
	Frame-Span	**31.95**	**43.94**	**35.63**	**41.16**	**49.45**	**44.18**	**53.52**	**71.43**	**59.34**
Trained on Multilingual	no vision	44.86	54.57	47.89	45.39	61.33	52.13	50.47	67.97	56.41
	Video-GIT	48.38	60.43	52.06	48.24	66.29	55.88	51.96	68.64	57.32
	Video-XCLIP	49.47	58.86	52.30	51.53	65.49	57.35	51.86	70.53	57.96
	Frame-Cos	50.47	61.37	53.76	49.11	70.80	58.18	52.66	69.97	58.16
	Frame-Span	**50.57**	**64.36**	**54.72**	**51.82**	**72.16**	**60.34**	**54.39**	**72.36**	**60.21**

Table 7: F1 scores of the anaphora resolution with video and frame features of gold mentions.

tions. According to the analysis of the true negatives, we observe that the decrease in performance is caused by zero anaphor detection. Multilingual training significantly enhances English mention detection, while only slight improvements are observed for Turkish and German. A similar effect of multilingual training on mention detection is illustrated in Figure 7, where outlier gold mentions (green dots) can be observed for German and Turkish, while English gold mentions are grouped together.

B.2 Multimodal with Language Experiments

Our language experiments address three distinct questions in Table 7. The first question investigates the effect of multimodal features on *cross-lingual* anaphora resolution. The second question examines *multilingual* anaphora resolution to assess the impact of multimodal features when both training and testing data are available for multiple languages. The last question pertains to *unseen* languages: how multimodal features impact the outcomes of anaphora resolution in languages that were not included in the training data.

Monolingual Monolingual results are obtained by training and testing the model in the same language. Using multimodal features with the Frame-Span method significantly enhances the results compared to using only language features (i.e., without the vision component) for both NP and zero anaphora resolution. For example, Turkish results demonstrate approximately a $\sim 4\%$ improvement for full anaphora resolution, around a $\sim 6\%$ improvement for zero anaphora resolution, and about a $\sim 2\%$ improvement for NP anaphora resolution. German and English monolingual results exhibit a similar improvement trend. Thus, it can be confidently asserted that multimodal features are effective for monolingual anaphora resolution, irrespective of the language family.

Multilingual Multilingual results are obtained by training and testing the model in Turkish, German, and English. The utilization of multimodal features outperforms language-only features across all evaluated cases, such as NPs, zero, and full anaphora resolution. The significant improvements associated with utilizing multimodal features are particularly evident in zero anaphora resolution, $\sim 10\%$ for German and Turkish and $\sim 5\%$ for English. For NP anaphora resolution, we observe approximately a $\sim 6\%$ improvement for both Turkish and German, and about a $\sim 4\%$ improvement for English. Notably, we do not observe significant improvement with multilingual features compared to monolingual experiments.

Unseen Languages The results pertain to testing on unseen languages that were not included

		Pronoun Omitted Turkish		
		NPs	Zero	Full
Trained on PO Turkish	no vision	44.05	61.75	51.38
	Frame-Cos	44.37	61.72	51.56
	Frame-Span	46.62	63.42	53.51
Trained on German	no vision	28.05	52.83	35.76
	Frame-Cos	30.51	56.01	39.38
	Frame-Span	34.83	56.28	42.44
Trained on English	no vision	18.65	12.25	16.20
	Frame-Cos	21.85	22.57	22.12
	Frame-Span	25.41	28.46	26.61
Trained on Multilingual	no vision	43.53	61.42	50.92
	Frame-Cos	44.71	61.91	51.75
	Frame-Span	45.42	66.06	53.86

Table 8: F1 scores of the anaphora resolution for Turkish test dataset with omitted pronouns. Multilingual dataset includes German, English and PO Turkish.

in the training process. For instance, the model is trained in the German dataset and tested on the Turkish dataset, or otherwise. By the results, multimodal features play a crucial role in enhancing the performance of anaphora resolution when applied to unseen languages, significantly improving the models' ability to generalize and effectively handle linguistic variations that were not present during the training phase for NP and zero anaphora resolution with the candidate, as in Table 3 and 4, and gold mentions as in Table 2 and 7.

B.3 Results of Visual Experiments

Video understanding studies use two approaches image-based encoding (Radford et al., 2021) with video frames, or video-based encodings (Ni et al., 2022; Wang et al., 2022) for video segments. Video-based encoding methods still use frames as inputs but focus on both the temporal and spatial features of the frames. We examine the performance of both methods for anaphora resolution. We compare the provided video features with our best-selected frames. In Table 7, we observe the features of best frames outperform the Video-XCLIP and Video-GIT features with Frame-Cos and Frame-Span methods for monolingual and multilingual experiments. We assume video features cause a similarity issue between the instruction. In Figure 6, two video segments are similar to each other when compared using the best frames, even though "beer" and "the meat" are different spans.

B.4 Results of Pronoun Omitted (PO) Turkish

The pronoun might be omitted in pro-drop languages when it is pragmatically or grammatically inferable from the context. Thus, we compare the cross-lingual and multilingual Turkish results with/out multimodal features between direct and pro-drop translations. The PO Turkish results are slightly reduced compared to those directly translated Turkish, multi-lingual, and unseen (Table 8) language experiments. The results of models trained with German and English drop drastically for zero anaphora, Table 8. Additionally, the PO Turkish-trained model yields inferior results compared to the direct translation approach.