

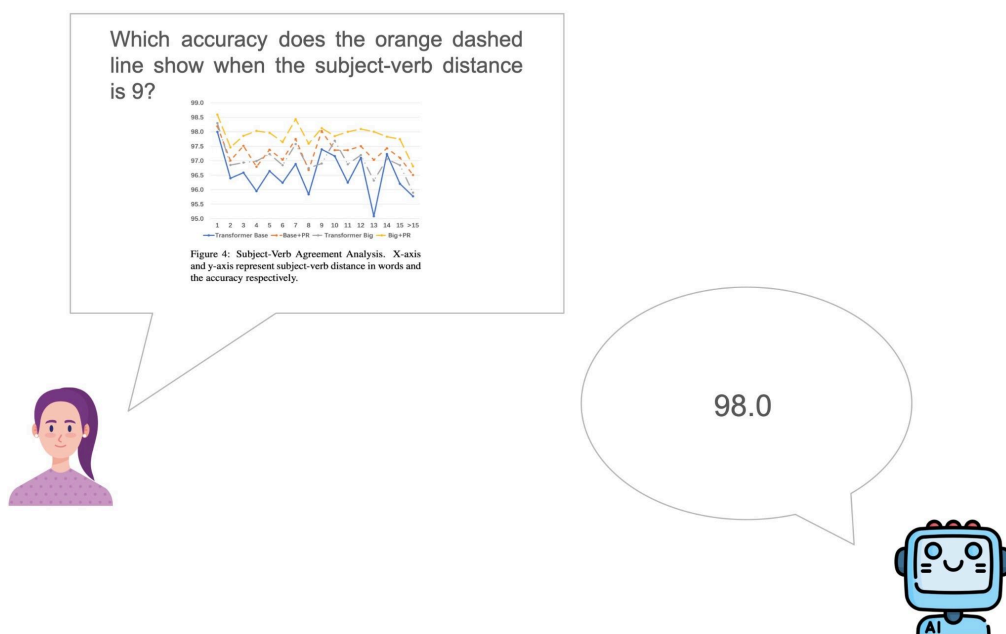


## Annotation Guidelines

# Introduction

## 1. Background

This annotation task involves creating question answering (QA) pairs and identifying types for figures extracted from scholarly publications (*scientific figures*). The data you are about to annotate was collected from Computer Science and Computational Linguistics papers available on [arXiv](#) and [ACL Anthology](#). The annotated corpus will be used in the Scientific Visual Question Answering (SciVQA) shared task run under the umbrella of [SDP workshop](#) at ACL 2025. In this challenge participants will develop systems able to take an image of a scientific figure, its caption, optionally additional metadata (e.g, figure type), and a natural language question as an input and output a natural language answer.



SciVQA corpus comprises 3000 figures and will be associated with 21000 QA pairs in total (see [Sec. 2](#)). This set of images will be equally divided between annotators and each figure will be annotated by one person. In what follows, [Sec. 2](#) provides the main concepts essential for the annotation procedure, [Sec. 3](#) contains the instructions on the annotation tasks, and [Sec. 4](#) explains how to access and use the annotation tool.

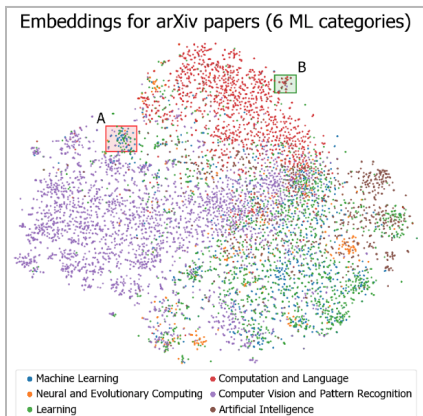
## 2. Main concepts

### a. Figure types

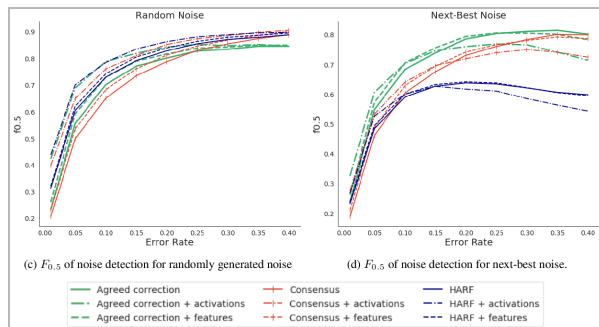
In the corpus under annotation, there are images with both *non-compound* and *compound* scientific figures. A non-compound figure contains **a single figure object** which cannot be

decomposed into multiple subfigures, while a compound figure involves **several (sub)figures** which can be separated and constitute individual figures. Here are the examples:

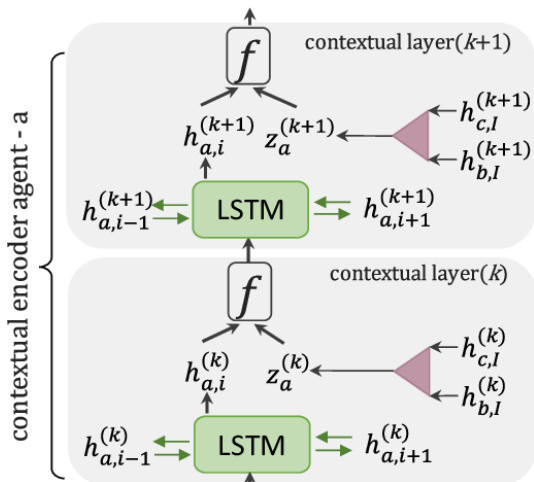
Non-compound Figure



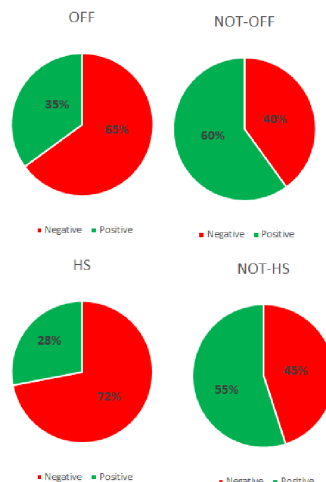
Compound Figure



Non-compound Figure



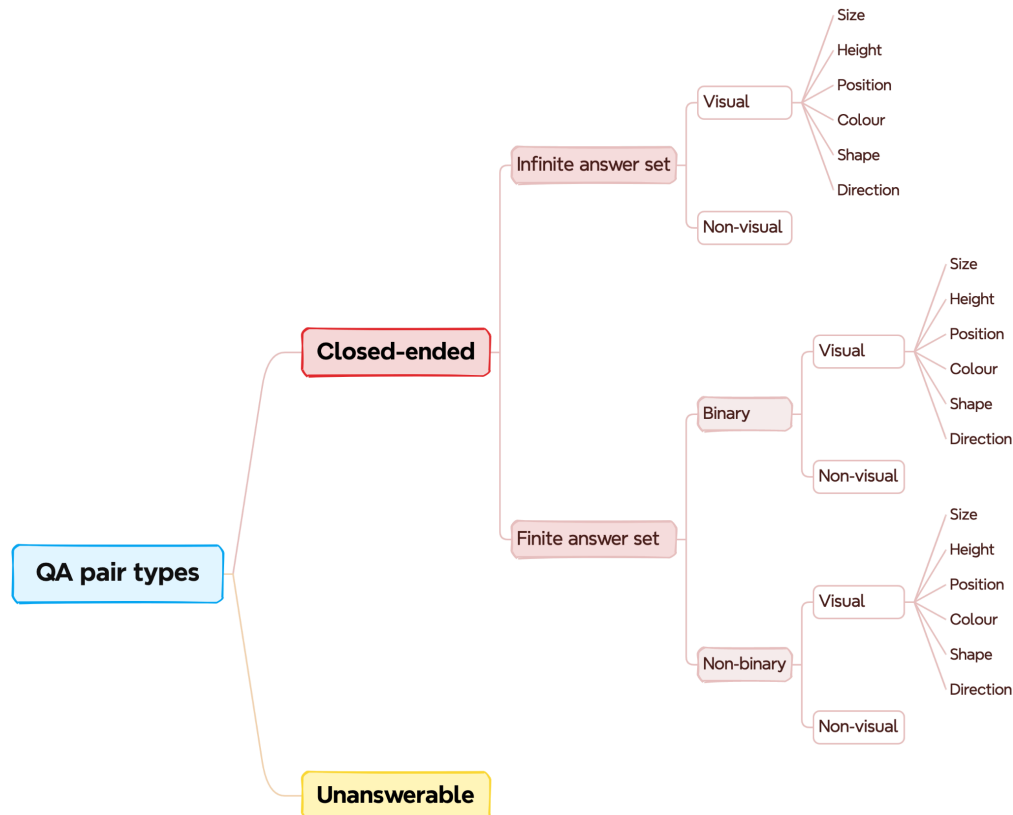
Compound Figure



Note that the bottom left figure is considered to be non-compound since all its elements are connected (e.g., arrow between the gray blocks, curly bracket related to both blocks) and splitting it into two separate figures is not possible. This is quite common for figures displaying system architectures. Additionally, figures can be classified into various types based on their visual representation. For instance, there are pie and bar charts, line graphs, scatter and box plots, etc. in the SciVQA dataset (see examples in [Appendix A](#)).

## b. QA pairs types

The images of the described types of figures will be annotated with QA pairs according to the schema below:



Here you can see that questions fall into two root classes: *closed-ended* and *unanswerable*. Overall, there are **seven** different question pair types which will be presented in the following. A closed-ended question implies that it is **possible to answer it based only on a given data source** (an image and a caption), i.e., no additional resources such as the main text of a publication, other documents/figures/tables, etc. is required. Thus, given the chart and caption below, one can ask the question "*What is the percentage of Twitter accounts related to Sports?*". In this case, the answer (9%) can be retrieved directly from the pie chart.

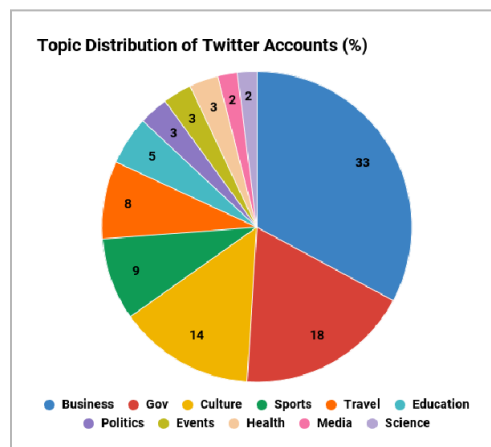


Figure 4: Distribution of accounts according to topic

On the contrary, an unanswerable question means that it is **not possible to infer an answer based solely on a given data source** (e.g., full paper text is required, values are not visible/missing, etc.). Thus, if we ask the question *"What is the exact number of Twitter accounts related to Sports?"*, one would need to have access to the additional information on the number of Twitter accounts (e.g., refer to the paper text or some other resources related to the study) since these details are not stated in the chart and its caption. Let's have a look at an example of an unanswerable question:

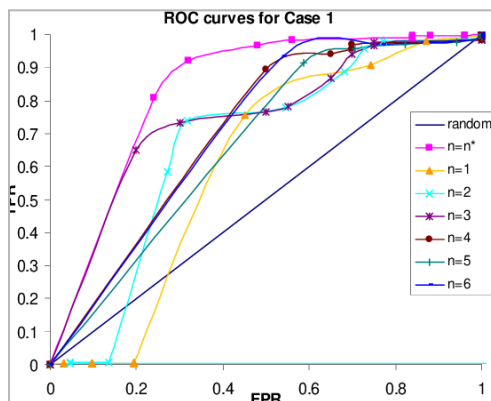


Figure 4. ROC curves from  $n^*$ -best and  $n$ -best.

The image of this line graph was not extracted properly from a paper PDF and as a result the name of the y-axis is missing. Therefore, if we ask *"What is the name of the y-axis?"*, it would be not possible to retrieve an answer from neither the figure image nor the caption. However, if we take as an example another figure (below), we can see that it is possible to ask *"What is the highest MP value achieved on the CUB dataset during training?"*. Since the MP values are visible on the y-axis and even though we cannot provide the exact number, we can indicate a range *"between 0.065 and 0.073"* or *"0.065-0.073"*. Here an unanswerable question could be *"What is the size of the CUB dataset used for the experiments?"* as this information cannot be inferred from the figure image or from both image and caption.

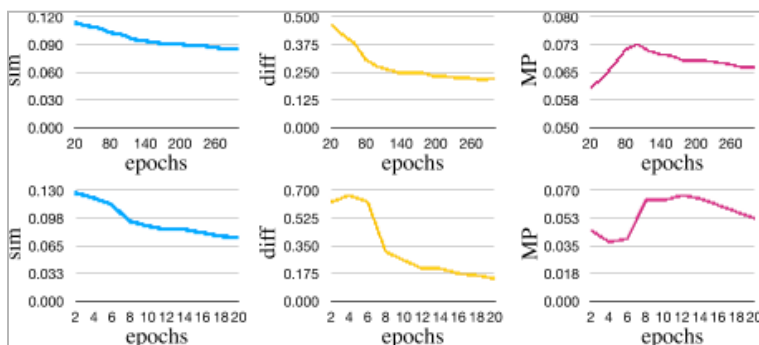


Figure 5: Text-image similarity (sim), L1 pixel difference (diff), and manipulative precision (MP) values at different epochs on the CUB (top) and COCO (bottom) datasets. We suggest to stop training the DCM module when the model gets the highest MP values shown in the last column.

But be careful as there could be tricky cases with piecewise-linear interpolation with QA pairs which you might consider to be answerable at the first glance. For instance, given the graph below, if we ask a question “What is the accuracy of the CNN with kernel size 4 when the feature map is 300?” it is in fact **unanswerable**. The reason is that there are only 3 data points provided in the graph, i.e.,  $x = 2.0$ ,  $x = 3.0$ , and  $x = 5.0$ , which are connected by a dashed line. There are no data points shown for  $x = 4$ . Thus, we cannot make any conclusions due to the lack of the information for a given datapoint.

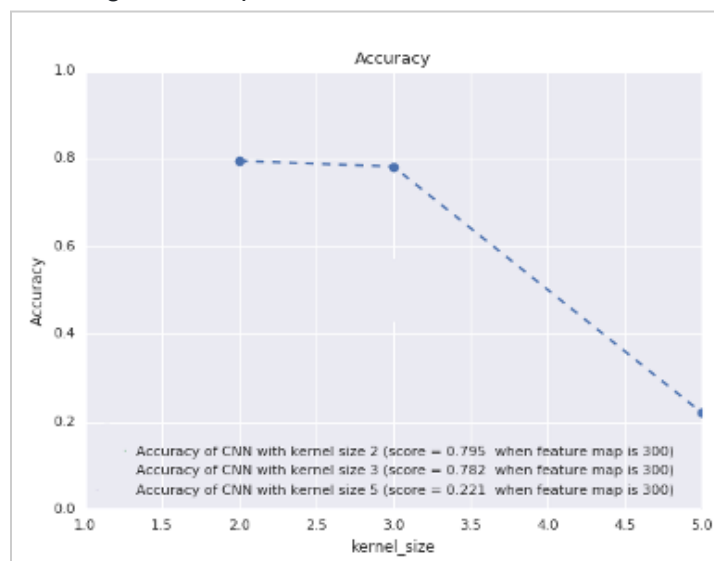


Figure 1: CNN accuracy for different kernel sizes when feature map is 300

So far so good! 🍌

Let's move to the second level in our QA pair types schema. Here the categorisation is based on the fact that for a given question  $Q$ , there exists a set  $S$  of all possible answers  $S = \{a_1, a_2, \dots, a_N\}$ , which can be either *infinite* or *finite*. As the name suggests, questions with an infinite  $S$  of answers simply **do not have any predefined answer options**, e.g., “What is the sum of  $Y$  and  $Z$ ?”. On the contrary, questions with a finite  $S$  of answers are associated with **a limited range of answer options**. Such QA pairs fall into two subcategories:

- **Binary** - require a “yes/no” or “true/false” answer. In this case, the answer options are limited by a question itself, e.g., “Is the percentage of positive tweets equal to 15%?”.
- **Non-binary** - require to choose from a set of  $M$  predefined answer options where one or more are correct. Here an  $S$  of all possible answer options is limited on purpose by those who define the question, e.g., “What is the maximum value of the green bar at the threshold equal to 10?” Answer options: “A: 5, B: 10, C: 300, D: None of the above”. In this annotation project,  $M = 4$ .

Each of the discussed above QA pair types can be *visual and non-visual*. Visual questions **address or incorporate information on visual attributes** of a figure, i.e., **shape, size, position, height, direction** or **colour**, e.g.,:

**Question:** “What is the minimum value of the **green line**?”

**Visual aspects:** Colour (green) and shape (line).

**Question:** “What is the difference in values between **the highest** and **the lowest bars**?”

**Visual aspects:** Height (highest/lowest) and shape (bars).

**Question:** “In the **bottom left figure**, what is the value of the **blue line** at an AL of 6?”

**Visual aspects:** Position (bottom left), colour (blue), and shape (line).

**Question:** “Which topic constitutes the **largest segment** in the **pie chart**?”

**Visual aspects:** Size (largest) and shape (pie chart).

**Question:** “In which direction does the output of the LSTM (**green box**) flow: **toward the** contextual layer  $k+1$  or **back to** layer  $k$ ?”

**Visual aspects:** Colour (green), shape (chart), and direction of objects (toward/back to).

Note that an answer would not necessarily contain visual information. Thus, visual QA pairs mention/contain information on the six visual attributes either in the question, answer or both. On the contrary, non-visual questions **do not involve any of the six visual aspects** of a figure defined in our schema, e.g., “What is the minimum value of  $X$ ?”, “What is the difference between the percentage of votes obtained for humor and non-humor tweets?”.

## QA types summary

Type	Short description
Closed-ended infinite answer set visual	<ul style="list-style-type: none"> <li>possible to answer based only on a given data source</li> <li><b>no</b> predefined answer options</li> <li>addresses visual aspects (colour, shape, height, etc.)</li> </ul>
Closed-ended infinite answer set non-visual	<ul style="list-style-type: none"> <li>possible to answer based only on a given data source</li> <li><b>no</b> predefined answer options</li> <li>does <b>not</b> address visual aspects (colour, shape, height, etc.)</li> </ul>

<b>Closed-ended finite answer set binary visual</b>	<ul style="list-style-type: none"> <li>possible to answer based only on a given data source</li> <li>limited answer options</li> <li>yes/no/true/false answer</li> <li>addresses visual aspects (colour, shape, height, etc.)</li> </ul>
<b>Closed-ended finite answer set binary non-visual</b>	<ul style="list-style-type: none"> <li>possible to answer based only on a given data source</li> <li>limited answer options</li> <li>yes/no/true/false answer</li> <li>does <b>not</b> address visual aspects (colour, shape, height, etc.)</li> </ul>
<b>Closed-ended finite answer set non-binary visual</b>	<ul style="list-style-type: none"> <li>possible to answer based only on a given data source</li> <li>limited answer options</li> <li>pre-defined answer set (A:... B:... C:....D:....)</li> <li>addresses visual aspects (colour, shape, height, etc.)</li> </ul>
<b>Closed-ended finite answer set non-binary non-visual</b>	<ul style="list-style-type: none"> <li>possible to answer based only on a given data source</li> <li>limited answer options</li> <li>pre-defined answer set (A:... B:... C:....D:....)</li> <li>does <b>not</b> address visual aspects (colour, shape, height, etc.)</li> </ul>
<b>Unanswerable</b>	<ul style="list-style-type: none"> <li>can <b>not</b> be answered due to the lack of information</li> </ul>

### 3. Tasks and instructions

The annotation project involves two main subtasks: [figure type classification](#) and [QA pairs validation](#). You will be assigned 500 images of the figures which must be annotated according to the [schedule](#). Additionally, there will be a possibility to access the figure source full paper PDF but **only in the edge cases**.

The annotation will be split into three phases:

- Phase 1.** Training of annotators (20 images);
- Phase 2.** Train set validation (380 images);
- Phase 3.** Test set validation (100 images).



Before diving into the tasks instructions, here are some general rules:

1. Read the guidelines **carefully**, do not proceed with the annotation without a good understanding of the main concepts and tasks.
2. It is advised to double check the results after validating each instance.
3. It is recommended to take **~3-5 min** breaks every **~3-5 images** and longer **~15 min** breaks every **~10-15 images**.
4. The deadlines are **strict**, please, stick to the schedule.
5. **The use of any models such as ChatGPT or Gemini to annotate the data is strictly prohibited!** This will violate the whole idea of manual validation.
6. If you are finished with the set of images before the deadline, let the supervisor know and start annotating the **next set**.
7. Do **not** access the projects assigned to other people.
8. In case of doubts or questions, contact the supervisor ([ekaterina.borisova@dfki.de](mailto:ekaterina.borisova@dfki.de)).

## Schedule

Number of images	Deadline
<b>Train set</b>	
20 images (training phase)	<b>17.02.25-21.02.25</b>
76 images (total 96)	<b>24.02.25-02.03.25</b>
76 images (total 172)	<b>03.03.25-09.03.25</b>
76 images (total 248)	<b>10.03.25-16.03.25</b>
76 images (total 324)	<b>17.03.25-23.03.25</b>
76 images (total 400)	<b>24.03.25- 27.03.25</b>
<b>Test set</b>	
100 images	<b>31.03.25-11.04.25</b>
<b>Human validation</b>	
20 images (testing the set up, gathering feedback)	<b>15.04.25-23.04.25</b>
80 images	<b>24.04.25-14.05.25</b>

## Figure types classification

Figures in SciVQA corpus were automatically categorised in two types, based on the two schemas discussed in [Sec. 2](#). For this, the [Gemini](#), a multimodal large language model, was used. Thus, figure types classification will involve two main tasks. First, you will be asked to verify the correctness of the *compound* and *non-compound* labels assigned to a set of images (see [Sec. 2](#)) as well as to confirm the *number of figures* present in each image. Second, for **some of the instances** you will be asked to validate whether those were correctly classified into one of the following categories: *bar chart*, *box plot*, *confusion matrix*, *line chart*, *pie chart*, *scatter plot*, *pareto chart*, *venn diagram*, *architecture diagram*, *neural network*, or *tree*. Please, see the examples of the listed charts in [Appendix A](#). Therefore, your job will be to approve or correct pre-assigned figure types.

## QA pairs validation

After completing the figure classification task, you will be asked to review the QA pairs pre-assigned to figures. In particular, all images are already associated with synthetic QA pairs generated by the [Gemini](#) model. There are **seven QA pairs per each image** which correspond to the seven types discussed in [Sec. 2](#). Let's have a look at the examples per each of the QA pair types based on the following non-compound line chart:

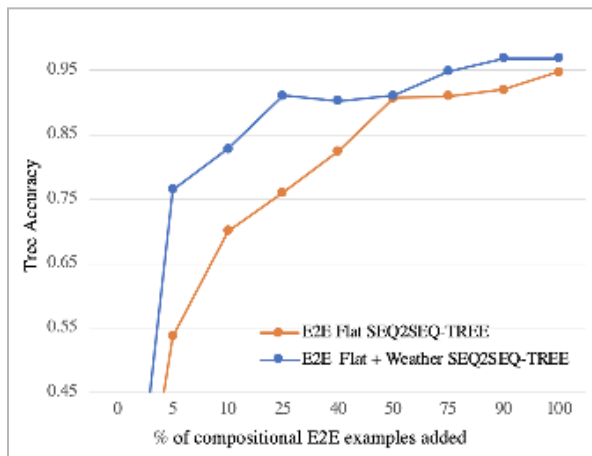


Figure 3: Performance of S2S-TREE models trained on E2E flat data, and flat E2E + full weather dataset, with a fraction of composition E2E.

1. **Closed-ended infinite answer set visual:**
  - **Question:** What is the approximate tree accuracy value of the orange line at 100% of compositional E2E examples added?
  - **Answer:** 0.95
2. **Closed-ended infinite answer set non-visual:**
  - **Question:** What is the percentage of compositional E2E examples added for E2E Flat data when the tree accuracy is equal to 90%?
  - **Answer:** Between 0.85-0.95.
3. **Closed-ended finite answer set binary visual:**
  - **Question:** Is the orange line consistently below the blue line on the graph?
  - **Answer:** Yes
4. **Closed-ended finite answer set binary non-visual:**
  - **Question:** Do the two models have the same performance at 90% of compositional E2E examples added?
  - **Answer:** No
5. **Closed-ended finite answer set non-binary visual:**
  - **Question:** Which line represents the performance of models trained on E2E Flat data?
  - Options:** A: The blue line, B: The orange line, C: The yellow line, D: Neither line
  - **Answer:** B
6. **Closed-ended finite answer set non-binary non-visual:**
  - **Question:** Using which training data resulted in the highest Tree Accuracy when 100% of the compositional E2E examples were added?
  - Options:** A: E2E Flat data, B: Compositional E2E, C: E2E Flat + full weather dataset, D: None of the above
  - **Answer:** C
7. **Unanswerable:**
  - **Question:** What is the size of the E2E Flat dataset?
  - **Answer:** It is not possible to answer this question based only on the provided data.

As can be seen from these examples (and as was mentioned in [Sec. 2](#)), **non-binary questions** are always provided with **four answer options** where one or more are correct.

**Your task will be:** given an image of a figure, its caption, and seven synthetic QA pairs with type information, check that QA pairs compile with a set of criteria listed below.

## Validation criteria

Indicate the QA pair as incorrect and modify it if it **violates** one of the following criteria or it **repeats/is very similar to** another QA pair associated with a given figure:

☒ **Questions are defined and answerable solely based on either a given figure or a given figure and its caption. But not based only on caption text (i.e., can be answered only given a caption text), there should always be information from the image included.** The only exception is an unanswerable QA pair type.

Note that QA pairs might require the general background knowledge in the field/topic, e.g., specific terminology. However, they must not require access to the context, i.e., full paper text or other additional resources. For instance, in case the question addresses the specific dataset used in the study, its size, models evaluated, etc., but this information is not provided in the caption and/or the figure but in the full paper text, such question is unanswerable.

✓ **Formulated questions correspond to the definition of a specified QA pair type.**

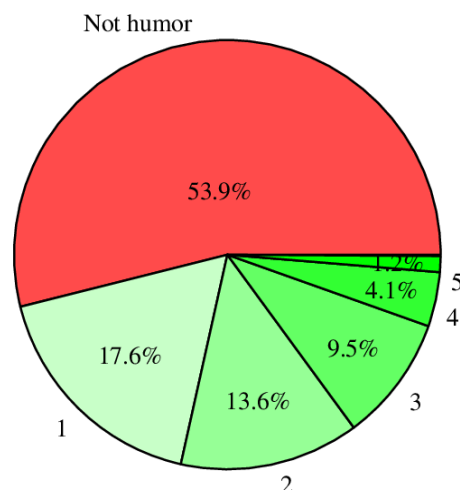


Figure 2: Distribution of votes in the final version of the corpus. The numbers 1 to 5 are the different scores the annotators could assign to the humorous tweets.

**QA type:** closed-ended infinite answer set visual

✓ **Question:** What is the sum of the red and green segments? **Answer:** 99.9%

✗ **Question:** How many tweets were found as not humorous? **Answer:** 1000 tweets

The QA pair on the left violates both the first and the second criteria. Namely, the information on the number of tweets is not provided neither in the image nor in the caption. Also the question does not address any of the visual attributes of the pie chart.

There could also be tricky cases where the question itself is binary but has multiple answer options (and some of them are binary):

**QA type:** closed-ended finite answer set non-binary non-visual



**Question:** Is the number of response vertices  $|R|$  fixed to 2?  
**Options:** A: True, B: False, C: Both, D: None

✓ **Non-binary questions are associated with four answer options.**

✓ **Question:** Which line represents the performance of models trained on E2E Flat data? **Options:** A: The blue line, B: The orange line, C: The yellow line, D: Neither line  
**Answer:** B

✗ **Question:** Which line represents the performance of models trained on E2E Flat data? **Options:** A: The blue line, B: The orange line  
**Answer:** B

✓ **Question does not call for the type classification.**



**Question:** Which of the following best describes the shape of the figure?  
**Options:** A: Histogram, B: Boxplot, C: Scatterplot, D: Line Graph

✓ **Answer options in non-binary questions are semantically diverse.**

✓ **Question:** Which of the following is NOT a vocabulary found in the Venn diagram? **Options:** A: WordNet with linguistic heuristics (WNling), B: UMLS, C: word2vec, D: SNOMED.

✗ **Question:** Which of the following is NOT a vocabulary found in the Venn diagram? **Options:** A: WordNet with linguistic heuristics (WNling), B: UMLS, C: word2vec, D: Word2Vec.

✓ **Answers are correct.**

✓ **Answers are rather short, i.e., do not contain any long explanations from a model including any extra information.**

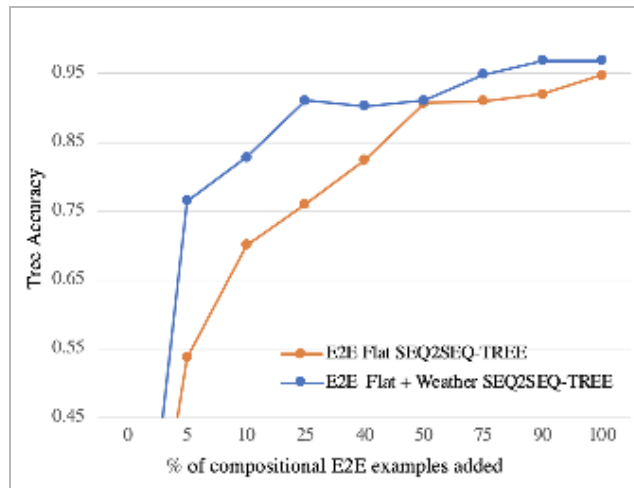


Figure 3: Performance of S2S-TREE models trained on E2E flat data, and flat E2E + full weather dataset, with a fraction of composition E2E.

**Question:** Is the orange line consistently below the blue line?

✓ **Answer:** Yes.

✓ **Answer:** Yes, the orange line is consistently below the blue line.

✗ **Answer:** No. Based on the provided graph, the orange line consistently falls above the blue line. This indicates that the "E2E Flat SEQ2SEQ-TREE" model consistently has a higher tree accuracy compared to the "E2E Flat + Weather SEQ2SEQ-TREE" model across all percentages of compositional E2E examples added.

The answer on the right is incorrect and includes extra information which makes the answer too long.

✓ Both questions and answers are grammatically, syntactically and semantically correct.

✓ Questions per image are not repetitive but as diverse as possible.

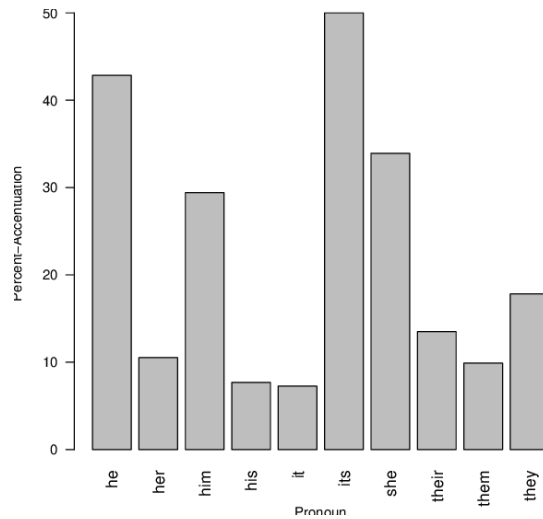


Figure 2: Variation between pronouns

**Question 1:** What is the approximate percent-accentuation of the pronoun 'it'?



**Question 2:** Which pronoun has the highest percentage-accentuation?

☒ The question is not ambiguous.

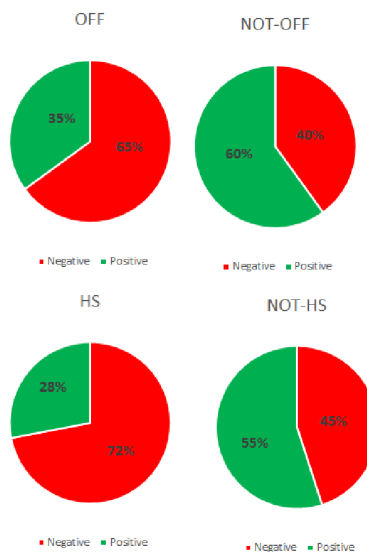


Figure 3: Distribution of Negative and Positive Tweets after applied AraNet on Shared-Task TRAIN Data

✗ **Question:** What is the distribution of positive tweets after applying AraNet on Shared-Task TRAIN Data?

✓ **Question:** What is the distribution of positive tweets after applying AraNet on Shared-Task TRAIN Data **for HS**?

✓ **Question:** What is the distribution of positive tweets after applying AraNet on Shared-Task TRAIN Data **in the bottom left chart**?

In the left example, it is not clear whether the question calls for the total percentage across all four pie chart or whether it refers to a specific graph.

✓ **Visual QA pairs explicitly address the visual attributes of a figure.**

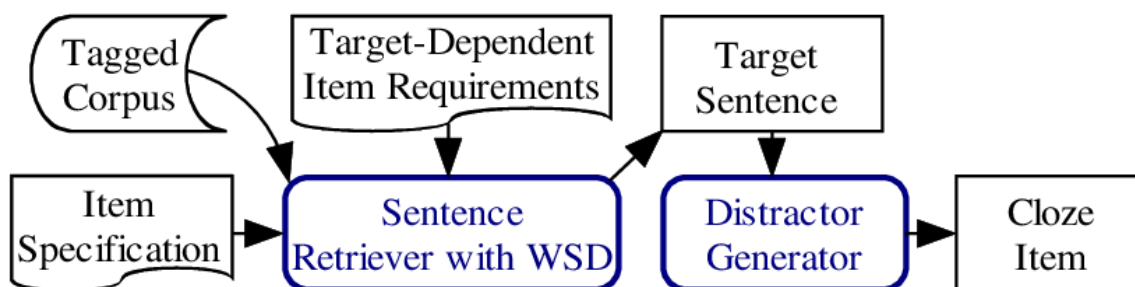


Figure 5. Main components of our cloze-item generator

✗ **Question:** What is the name of the component that receives input from both the Target Sentence and the Item Specification?

✓ **Question:** Is the Target Sentence connected to the Distractor Generator by an arrow?

## Correction of QA pairs

Additionally, there are **several rules on how to handle the correction of QA pairs**:

1. If it is **not possible at all** to create a specific QA pair type(s), the output should be **None**. **Except the case**, when a question is based solely on caption and it is **not** possible to ask it based on **both** an image and a caption. Then, you should leave the



question as it is (or if relevant modify, e.g., fix grammar mistakes, etc.) and put a note stating “**question is based solely on caption**”.

2. If there is an option to add **additional visual attribute(s)** to a given visual QA pair, do that. For instance, if a question addresses only the shape of an object (e.g., line, box, bar, dashed line, circles, etc.) but you also see that colour or any other information can be integrated, update the question accordingly.
3. If a synthetic QA pair addresses only **one subfigure** in a compound figure but could address **several subfigures** with minor corrections, reformulate it into a **cross-figure question**.
4. If **all binary QA pairs** in a given instance have a “**yes**” answer, **modify one** of them to bring diversity. Thus, we will reduce bias and imbalance towards questions towards this answer option.
5. If the question seems unclear or complex, first research unfamiliar terms. If it remains difficult or possibly incorrect, refine it **without oversimplifying**. E.g., changing “*Which of the following plots shows a similar trend in the 'Mp' values at different epochs for the CUB and COCO datasets?*” to “*Which of the following plots have a pink line?*” oversimplifies the question. In the correct version, we lose the numerical reasoning aspect and this is **not** what we want. The aim is to test all the abilities of a given model by challenging it!
6. If a question calls for a specific value but it is **not possible** to retrieve it, the range should be specified or you can add the word “approximately”. In case the latter is not possible, the question is considered to be unanswerable and should be tagged as such.

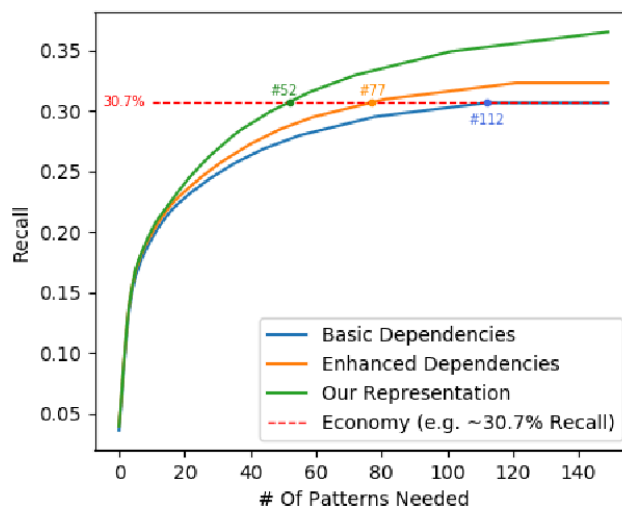


Figure 3: Economy comparison: Recall vs number of patterns, for the different representations.

**Question:** What is the recall value for the blue line when the number of patterns needed is equal to 80?

✗ **Answer:** 0.28.

✓ **Answer:** 0.25-0.30.

✓ **Answer:** between 0.25 and 0.30.

7. In case the answer is **incorrect** due to a mistake in the question (e.g., wrong colour, value, name, etc. is addressed), the correction can be done to **either** the question **or** the answer.

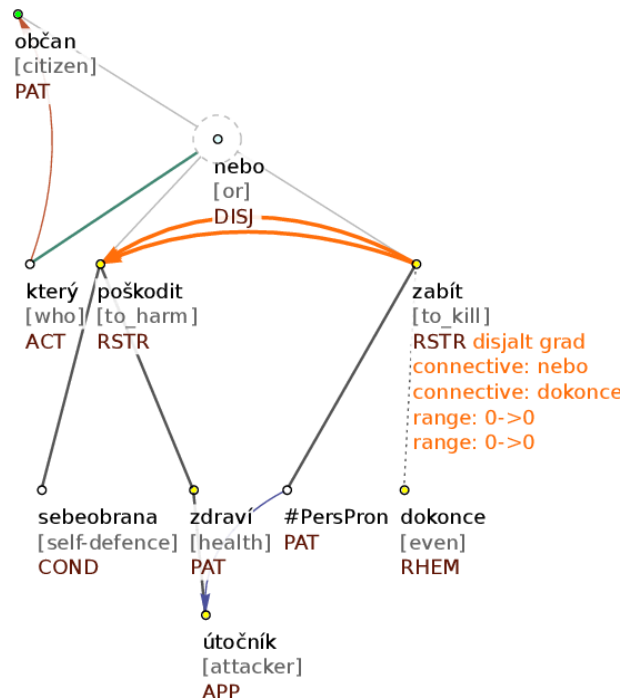


Figure 1: Annotation of discourse relations in PDiT 2.0. The relations are represented by two orange arrows connecting roots of the arguments. Information about the discourse types and connectives is given at the starting node of the relations.

**Question:** Is there a blue arrow pointing towards a node labeled "zdraví"?

**Answer:** Yes

In this example the model was actually referring to the purple line but made a colour recognition error. Here it is possible to either correct "a blue arrow" into "a purple arrow" or modify the

answer to “No”. The choice might depend on how the initial question is formulated. In this example, the better solution might be to edit the colour reference in the question.

8. If there are **multiple colour shades**, stick to the **standard color names** (i.e., blue, red, orange, black, etc). You **can also use dark/light indicators** but do **not** use the specific shades as different palettes exist and the names vary across them.
9. In case a **compound figure contains various types** of diagrams/charts, e.g., 2 line graphs and 1 bar chart, choose **several** respective tags from the list.

### Remember

You can refer to the full paper text only if a given chart is too complex and you are not sure if a question/an answer is correctly defined.

Do not forget to always check the caption for a given figure since a model quite often relied on it.

## Human validation

You will be assigned a set of **120** figure images from the annotated and validated (in the previous phase) SciVQA **test set**. The task is given an image of a figure, its caption, type, and a question, to provide an answer. As you already know, there are 7 questions per instance.

### Instructions:

1. Go to the project called “yourname yoursurname-Human Eval - Test Split” . Please, refer to the [schedule](#) for the specific phases and deadlines.
2. The free form answers must be rather **short**, e.g., if a question calls for a specific value the answer should be just a number instead of a full sentence, **Q:** “*What is the percentage of tweets that are positive?*”, **A:** “12%”.
3. If it is not possible to answer a question based on a given data source (image and caption or one of those), please state the following in the answer field: **It is not possible to answer this question based only on the provided data**. It is the unanswerable type of question from the taxonomy.
4. In case you think there is no correct answer (for multiple choice), select the “**I don’t know**” box. Since for this type of question there should always be 1 or more correct answers. It would be helpful if you also leave a note stating why there are no correct

answers (e.g., colours/values cannot be seen) or/and what the correct answer is (if you know).

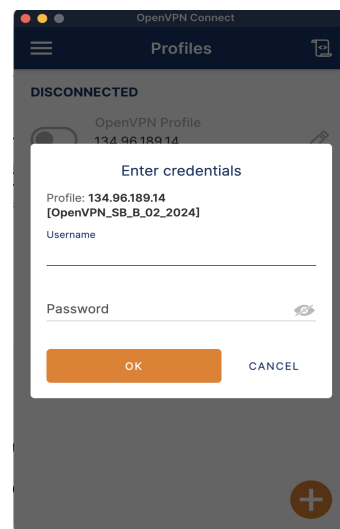
5. In case you cannot answer the question at all, e.g., because you cannot understand it, maybe the phrasing is confusing, simply select the “**I don’t know**” box. And please indicate a reason in addition to that in the Notes field. This point and also 4 are different from point 3 since the reason is other than the lack of information in the given image and caption.
6. **Important:** It is not allowed to search for an original paper PDF to gain additional information or to prompt models like GPT-4.

## 4.Annotation tool

### a. Access the project

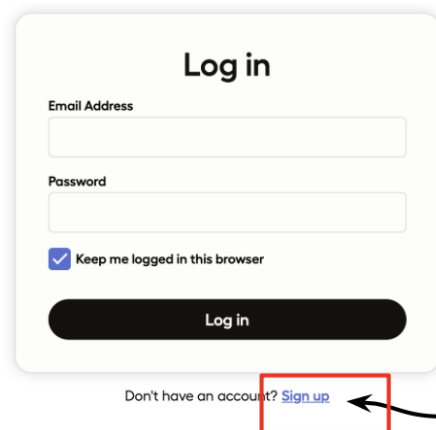
In this project, [Label Studio](#) is used as an annotation tool. To access it and the project assigned to you, follow the steps below:

1. Connect to the VPN with your DFKI credentials:



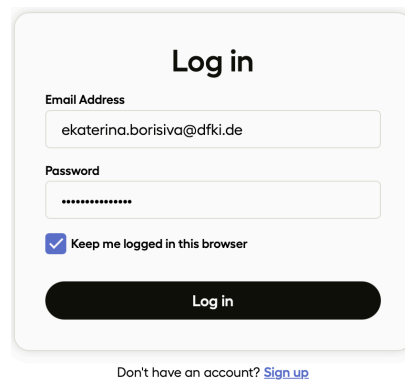
2. Access the Label Studio instance via: [\[link\]](#).

3. Create an account:



The image shows a login form titled "Log in". It contains two input fields: "Email Address" and "Password". Below the "Password" field is a checkbox labeled "Keep me logged in this browser" which is checked. At the bottom of the form is a black button labeled "Log in". Below the form, there is a link that says "Don't have an account? [Sign up](#)". The "Sign up" link is highlighted with a red rectangle, and a black arrow points to it from the right.

4. Log in:



The image shows the same login form as above, but with pre-filled data. The "Email Address" field contains "ekaterina.borisiva@dfki.de" and the "Password" field contains "\*\*\*\*\*". The "Keep me logged in this browser" checkbox is still checked. The "Log in" button is at the bottom. Below the form, the link "Don't have an account? [Sign up](#)" is visible.

5. Choose the project with your name and the relevant data split.

## b. General functionalities

1. You will always have 3 options:

- a. Indicate an annotation as being Correct
- b. Indicate an annotation as being Incorrect and Edit it.
- c. Add Notes in case you notice some general issues with an image /annotations (e.g., “image has poor quality, values are not readable”) or you have some minor comments (e.g., “modified the question to introduce diversity”).

But you are advised to contact the project lead directly instead.

**The image is: Non-compound**

☒ Correct<sup>[1]</sup> ☐ Incorrect. Edit<sup>[2]</sup> ☐ Add Notes<sup>[3]</sup>

**The image is: Compound**

☐ Correct<sup>[1]</sup> ☒ Incorrect. Edit<sup>[2]</sup> ☐ Add Notes<sup>[3]</sup>

☒ Non-compound<sup>[4]</sup> ☐ Compound<sup>[5]</sup>

2. Structure of the displayed content for QA pairs is straightforward: Type, Question (in yellow), Answer (in blue).
3. There is always an option to correct both a question and an answer.

But make sure that you actually modified one of those when you click Incorrect.Edit. Otherwise, do not forget to switch back to the Correct option in case you accidentally clicked on Incorrect.Edit.

Do not forget to click “Update”.

**Question 1**


Question Type: closed-ended infinite answer set visual

How many vertical rectangles are in the oval labeled "Lang-2"?


Answer: 5

☐ Correct<sup>[b]</sup> ☒ Incorrect. Edit<sup>[y]</sup> ☐ Add Notes<sup>[l]</sup>

How many vertical rectangles are in the oval labeled "La

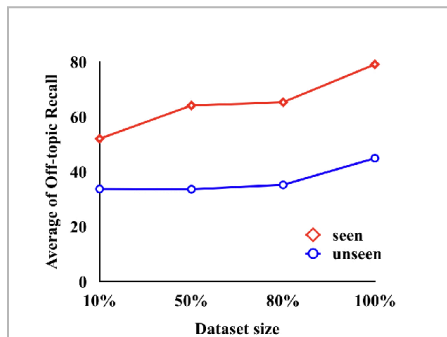
How many vertical rectangles are in the oval labeled "Lang-2"? 

5

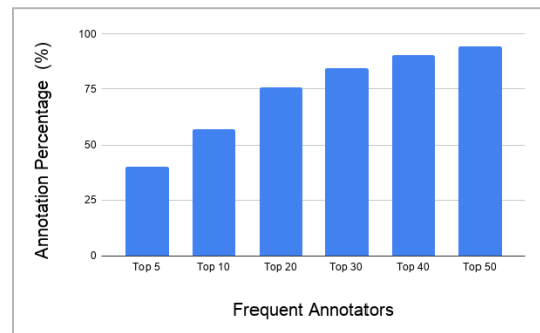
4 

## Appendix A

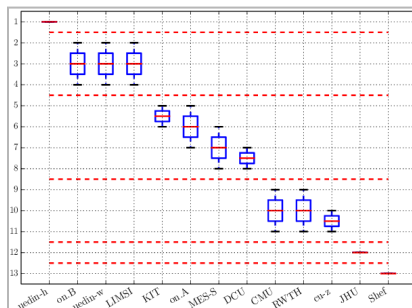
Examples of figure types for [Figure classification task](#) in annotation Phase I.



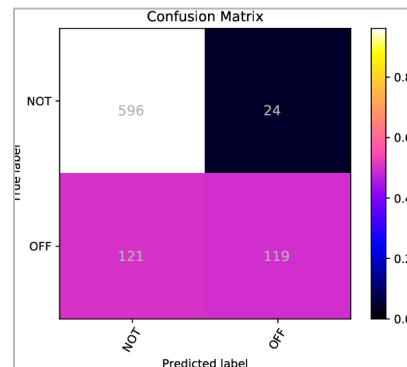
Line chart



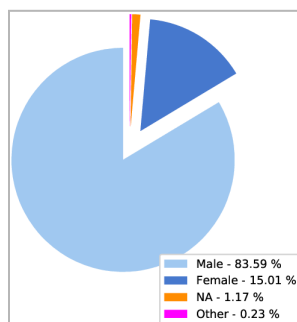
Bar chart



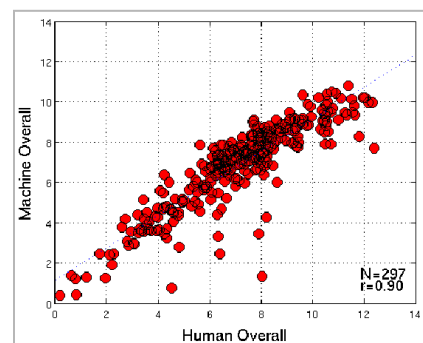
Box plot



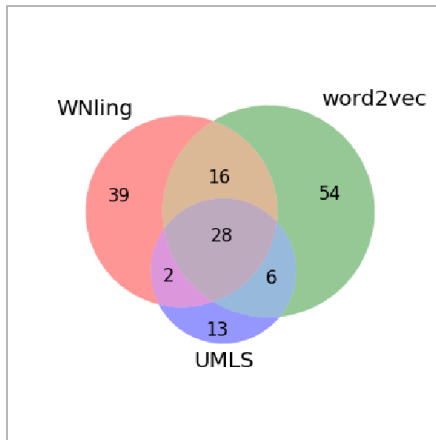
Confusion matrix



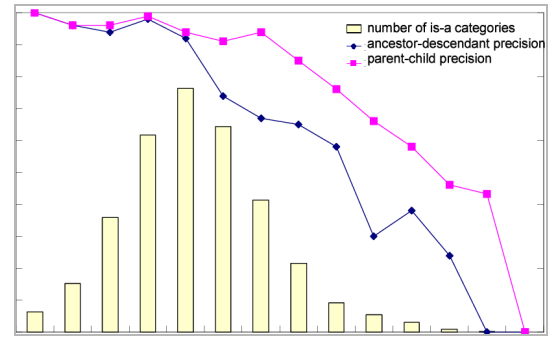
Pie chart



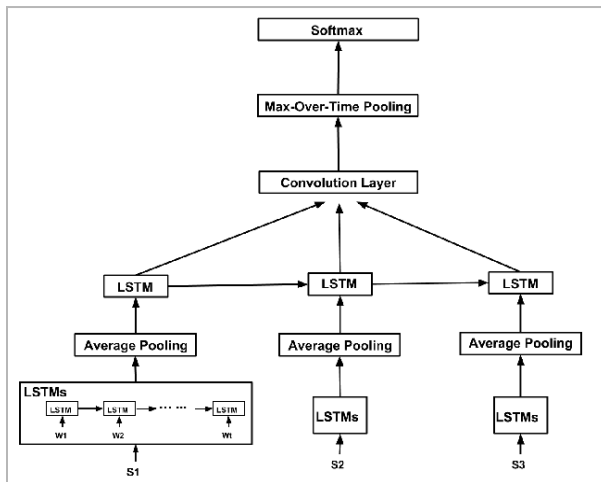
Scatter plot



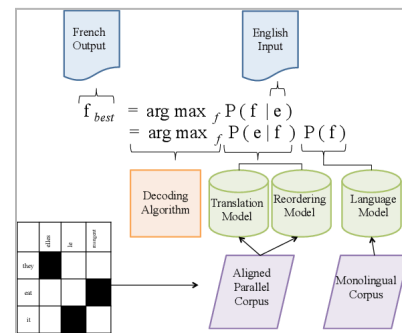
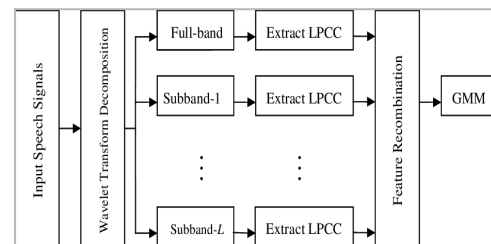
Venn diagram



Pareto chart

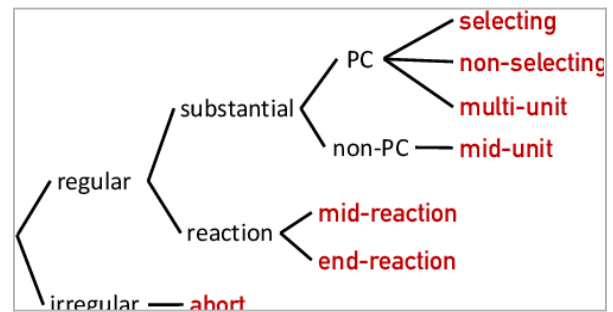
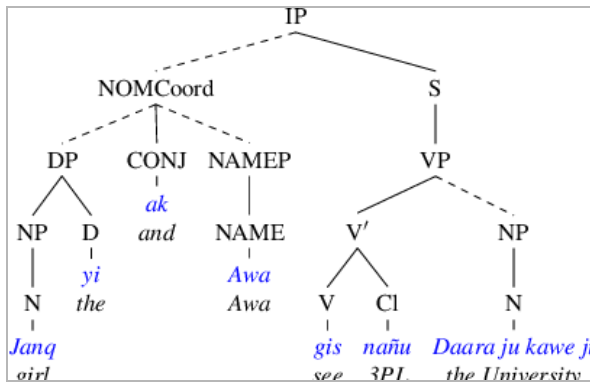


Neural networks

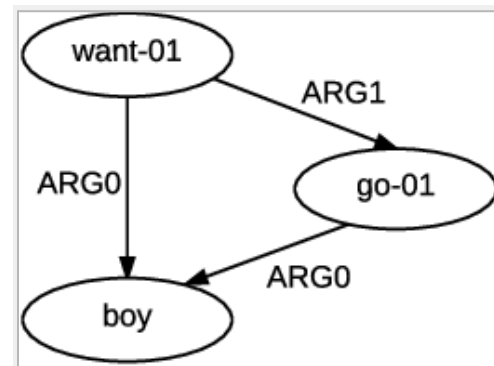
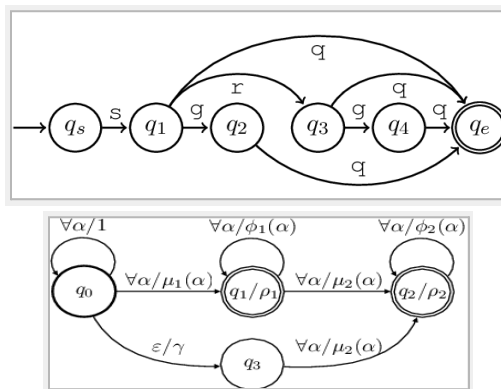


Architecture diagram





Trees



Graph