

Annotation Guidelines

Edits

Highlights in a specific color means when the edits were made:

07 Jun 2023

15 jun 2023

27 jun 2023

Goal/Task

In this annotation project, we are interested in knowing what the topic and genre is of a sentence and whether we humans can identify these. For Topics, we make use of the Dewey Decimal Classification (DDC) system. For genres, we make use of the genres provided in the Georgetown University Multilayer Corpus (GUM) corpus. The goal is to put the sentence/paragraph at hand into the most probable class (determined by you).

Genre has a *one-layer* annotation scheme, while **Topic** has a *two-layer* annotation scheme, which we will refer to as L1 and L2. We want to annotate for all three. There is an option for "Not Sure" (abbreviated to "NS"). This is when you feel that the label for the sentence is not present in the options. In addition, feel free to add any notes for clarification (e.g., clarify your choice or something else).

Preliminaries

Below we give an introduction to the topics and genre labels of this annotation project. It takes around 15-20 minutes to read. *Note that you don't have to remember the label numbers.* This introduction is to make you aware of the definition of the classes. All the labels are present in the annotation spreadsheet

Introduction Genres

We make use of the text types (genres) in the GUM corpus. These genres do not have a specific number like the topics above. Therefore we simply enumerate them. The genres are the following:

1. **Academic**
2. **Bio**
3. **Conversation**
4. **Fiction**
5. **Interview**
6. **News**

7. **Speech**
8. **Textbook**
9. **Vlog**
10. **Voyage**
11. **Whow**

Brief explanation of the genre classes:

1. **Academic** (writing) is nonfiction writing adhering to academic standards and disciplines. It includes research reports, monographs, and undergraduate versions. It uses a formal style, references other academic work, and employs consistent rhetorical techniques to define scope, situate in research, and make new contributions.
2. A **biography** is a detailed description of a person's life. It involves more than just basic facts like education, work, relationships, and death; it portrays a person's experience of these life events. Unlike a profile or curriculum vitae (résumé), a biography presents a subject's life story, highlighting various aspects of their life, including intimate details of experience, and may include an analysis of the subject's personality. Biographical works are usually non-fiction, but fiction can also be used to portray a person's life. One in-depth form of biographical coverage is called legacy writing. Works in diverse media, from literature to film, form the genre known as biography. An authorized biography is written with the permission, cooperation, and at times, participation of a subject or a subject's heirs. An autobiography is written by the person themselves, sometimes with the assistance of a collaborator or ghostwriter.
3. **Conversation**: naturally occurring spoken interaction. Represents a wide variety of people of different regional origins, ages, occupations, genders, and ethnic and social backgrounds. The predominant form of language use represented is face-to-face conversation, but also documents many other ways that people use language in their everyday lives: telephone conversations, card games, food preparation, on-the-job talk, classroom lectures, sermons, story-telling, town hall meetings, tour-guide spiels, and more.
4. **Fiction** refers to creative works, particularly narrative works, that depict imaginary individuals, events, or places. These portrayals deviate from history, fact, or plausibility. In our data, fiction pertains to written narratives like novels, novellas, and short stories.
5. An **interview** is a structured conversation where one person asks questions and another person answers them. It can be a one-on-one conversation between an interviewer and an interviewee. The information shared during the interview can be used or shared with others.
6. **News** is information about current events, shared through various media like word of mouth, printing, broadcasting, electronic communication, and witness testimonies. It covers topics such as war, government, politics, education, health, environment, economy, business, fashion, entertainment, sports, and unusual events. Government announcements and technological advancements have accelerated news dissemination and influenced its content.

7. A (political) **speech** is a public address given by a political figure or a candidate for public office, usually with the aim of persuading or mobilizing an audience to support their ideas, policies, or campaigns. Political speeches are an essential tool for politicians to communicate their vision, articulate their positions, and connect with voters or constituents.
8. A **textbook** is a book containing a comprehensive compilation of content in a branch of study with the intention of explaining it. Textbooks are produced to meet the needs of educators, usually at educational institutions. Schoolbooks are textbooks and other books used in schools. Today, many textbooks are published in both print and digital formats.
9. A **vlog**, also known as a **video blog** or **video log**, is a form of blog for which the medium is video. The dataset contains transcripts of the speech occurring in the video.
10. A **travel/voyage** guide is a **wiki** providing information for visitors or tourists about a particular place. It typically includes details about attractions, lodging, dining, transportation, and activities. It may also contain maps, historical facts, and cultural insights. Guide **wikis** cater to various travel preferences, such as adventure, relaxation, budget, or specific interests like LGBTQ+ travel or dietary needs.
11. A Wikihow **how-to (whow)** guide is an instructional document that offers step-by-step guidance on accomplishing a specific task or reaching a particular goal. It aims to assist individuals in learning and comprehending the process involved in successfully completing the task. These guides are typically written in a clear and concise manner, simplifying complex processes into manageable steps. They often include detailed explanations, diagrams, illustrations, or examples to enhance understanding. How-to guides cover various topics, such as technical tasks, practical skills, creative endeavors, troubleshooting, and more.

Introduction Topics

The DDC system is a widely used library classification system developed by Melvil Dewey in the late 19th century. The DDC is based on the principle of dividing knowledge (in our case sentences) into ten main classes, each identified by a three-digit number; **we only focus on the first two.**

1. The ten main classes in the Dewey Decimal Classification system are as follows:

000 Computer science, information & general works
100 Philosophy & psychology
200 Religion
300 Social sciences
400 Language
500 Science
600 Technology
700 Arts & recreation
800 Literature
900 History & geography

These higher level classes belong to L1 in the annotation spreadsheet, and we added the NO-TOPIC label (see description below)

2. Each main class is further divided into subclasses using additional digits (10s). For example, in the 500s (natural sciences and mathematics), you'll find 510 for mathematics, 520 for astronomy, 530 for physics, and so on. The system allows for more specific classification of books and materials based on their subject matter.

See the following page:

<https://www.oclc.org/content/dam/oclc/dewey/ddc23-summaries.pdf>

This page separates the ten classes above into more finer-grained classes. There is not an explanation for each of them, but usually the name of the label encapsulates the subclass already. Note that the subclasses overwrite the main classes (so you can't pick 400 and 510, then you'd have to change 510 to 500).

These subclasses belong to L2 in the annotation spreadsheet.

Note that for each fine-grained class we deem the main number/code (e.g., 100, 200, 300) in L2 as the No-topic/Other category. The "Other" class can only be chosen in the fine-grained label classes (L2). Choosing this means that you believe that the current sentence belongs to a specific class. But the label is not present.

The Dewey Decimal Classification system is used in many libraries around the world to organize their collections and make it easier for users to locate resources. It provides a systematic way of arranging materials and enables efficient browsing and retrieval of information based on subject areas.

Brief explanation of the topic classes (L1)

- **000 Computer science, information & general works** is the most general class and is used for works not limited to any one specific discipline, e.g., encyclopedias, newspapers, general periodicals. This class is also used for certain specialized disciplines that deal with knowledge and information, e.g., computer science, library and information science, journalism. Each of the other main classes (100-900) comprises a major discipline or group of related disciplines. **Note that in our experiments, we do not consider this a miscellaneous category, we have "No-topic" for this.**
- **100 Philosophy & psychology** covers philosophy, parapsychology and occultism, and psychology.
- **200 Religion** is devoted to religion.
- **300 Social sciences** covers the social sciences. Class 300 includes sociology, anthropology, statistics, political science, economics, law, public administration, social problems and services, education, commerce, communications, transportation, and customs.
- **400 Language** comprises language, linguistics, and specific languages. Literature, which is arranged by language, is found in 800.

- **500 Science** is devoted to the natural sciences and mathematics.
- **600 Technology** is technology.
- **700 Arts & recreation** covers the arts: art in general, fine and decorative arts, music, and the performing arts. Recreation, including sports and games, is also classed in 700.
- **800 Literature** covers literature, and includes rhetoric, prose, poetry, drama, etc. Folk literature is classed with customs in 300.
- **900 History & geography** is devoted primarily to history and geography. A history of a specific subject is classed with the subject.
- **No topic:** For cases where the topic can not be determined, or even guessed. For example for utterances that contain no natural language or do not have enough context.

FAQ

1. Should the colors of L1 and L2 in the annotation spreadsheet match?

Yes, apart from that the colours should match, the first number of the class to which the sentence belongs should also match.

For example, a sentence that belongs to Arts (700), is restricted to anything in the 700 class, e.g., a painting (750).

2. If a sentence has a clear topic in general, but the L2 category does not match, how do we annotate?

The fine-grained (L2) topics have the priority, and since they have to match you adjust the main topic accordingly.

3. Does my choice of Topic depend on the Genre or vice versa?

No, by default, annotating for genre and topic should be a separate task and should not influence each other.

4. How do we distinguish between something that is in the No-topic (or Others) class and NS ("not sure")?

Use the "others" category when you believe the current instance to belong to a class which is not in the listed ones. Mark your choice with "NS" when you have a guess, but you are not confident about it (e.g., because the instance is very short, or you are not familiar with the genre/topic)

If you are able to find L1, but none of the labels fit for the sentence in L2, you should choose "Other" (e.g, 000, 100, 200, etc.) in the same colour (class) of L2. The "Other" class can only be chosen in the fine-grained label classes (L2). Choosing this

means that you believe that the current sentence belongs to a specific class. But the label is not present. Otherwise, mark your best guess with "NS".

5. Is it better to label a sentence as "NO-TOPIC" if there is not a clear label associated with it or are we encouraged to take a guess?

You are encouraged to take a guess. However, for cases where you have no preference for any of the labels (i.e. a wild guess), label it as NO-TOPIC.

6. There is already another "Other" class in Religion/Language (e.g., 290 Other religion).

Good catch, imagine this situation. Let's say the sentence is talking about Buddhism. This falls under 290, because we're talking about another religion. However, if the sentence is "vaguely" talking about religion and doesn't fit within any of the labels, then choose 200 (Other).

7. Where do ads/exam questions fit?

In whichever of the genres you would expect to come across advertisements/exam questions. However, note that the data is scraped from the main information channel of source (i.e., advertisements next to a news text or before a vlog are not included).

8. Can we use external resources?

External resources are allowed, but do not look up the literal sentence 😊.


9. How to pick topics (L1/L2) for fiction (genre)?

Note that the genre and topic tasks should be seen as **distinct tasks**. So, the genre fiction should not automatically lead to a literature topic label (unless the fiction work is about literature).

10. Some utterances seem to be taken from the same text; do we have to give them the same label, or take the contexts into account?

No, each utterance should be judged independently.

Note for L3:

- For each L2, there is a finer-grained class namely L3. These numbers go in the thousands. Now, try to pick the most likely thousands' topic:
 - You will have to refer to the PDF ( L3-1000.pdf) for the right classes.
 - Please write the **class number** in the spreadsheet cell. There is no dropdown menu.
- The "no-topic" option still exists. Use "NT";

- You should pick the fine-grained L3 topic that best fits the utterance. This time you don't have to match the L1-L2 categories, but we ask you to **NOT** update your previous L1-L2 annotations, and just annotate L3 independently.