Annotation Guideline: Variable Detection and Linking

Version 0.3

Background

In social science literature, survey variables play an important role in the discussion and analysis of observed social phenomenon. As such, identifying where and which variables are mentioned in free text is vital for analyses. Currently, there is a lack of freely available data sets for the development and evaluation of tools for identifying variables, which motivates the current annotation task for Variable Detection and Linking. This annotation guideline has been written in order to assist annotators in the given task and provide users of the data, and derived models thereof, an understanding of how the data was collected.

Terminology

In order to promote a shared vocabulary, we use the following terminology for the remainder of this document:

- *Survey data set* - a questionnaire consisting of multiple question and answer choice pairs, called items or *survey variables*.
- *Survey variable* - an item from a *survey data set*.
- *Survey variable text* - the actual text of an item from a survey data set. Each variable contains different fields, such as a question and multiple answer choices.
- *Survey variable mention/reference* - an in-text mention of an item from a survey data set. It can be in the form of a (partial) quotation, a paraphrase, a negative polarity item, lexical inference, or different variable fields (examples can be found in the Appendix).
- *Annotated instance* (or just *instance*) - a piece of text containing one or multiple survey variable mentions and the linked survey variables.

Task Description

The goal of the annotation task is to create annotations in social science publications (or documents), which provide direct links to the corresponding survey variables that the authors of the publications are referring to. Given a document and a list of survey variables, an annotator is asked to read the full document and highlight all complete sentences that contain a survey variable mention. The annotation can thus be divided into three tasks:

- **Task 1** - Paragraph Detection:
  - Type: Paragraph-level
  - Description: For each document, identify and mark complete paragraphs that contain at least one sentence containing a variable mention.
- **Task 2** - Variable Detection:
  - Type: Sentence-level

  - Description: For each document, identify and mark all full sentences that contain at least one survey variable mention. Consecutive sentences that contain survey variable mentions should each be annotated.

- **Task 3** - Variable Linking:
  - Type: Sentence-level
  - Description: For each sentence containing a survey variable mention, select all relevant survey variables from the provided list.

The annotation has two phases. In *phase 1*, annotators annotate documents to get familiar with the annotation format and tools. The documents for *phase 1* are shorter than those for *phase 2*. The annotation is run in multiple rounds. Each round contains a set of documents that need to be annotated completely within a given timeframe.

Data

The data set for annotation consists of a diverse set of social science publications in both English and German. The publications range in size and topic. Certain publications may contain more survey variable mentions, while others very few or none at all.

Annotation Procedure

Step 0: Prepare environment

Log into the INCEpTION environment and the external recommendation tool. Open the "annotation_guideline.pdf", which can be found by clicking the "Guidelines" button (book icon). During *phase 1*, you are also provided with a survey variable catalog, called "survey_variables.pdf" (also found under "Guidelines"). During *phase 2*, you are provided with the catalog and an additional search tool, which are both integrated into the recommendation tool.

Step 1: Open document

View the current annotation project and open the next document in the list (which has not already been fully annotated) and open it in the PDF-view.

Step 2: Get an overview

Read the title and abstract (if available) to determine the topic of the text.

Step 3: Read the document

Read the document paragraph-by-paragraph (including the abstract). Identify, whether a paragraph contains a variable or not. Mark the entire paragraph using the "Paragraphs" layer (selected from the dropdown on the right sidebar). If a paragraph spans across pages, simply mark as many consecutive sentences as allowed by INCEpTION. Mark all remaining sentences that belong to the paragraph on the following page.

Step 4: Identify variable mentions

For paragraphs that contain a variable, identify and mark all sentences that mention a variable using the "Survey Variables" layer (selected from the dropdown on the right sidebar). Sentences may contain (part of) the variable text literally, or express the semantic content of the variable in other words, or be narrower/broader. Examples of different types of references to variables can be found in the Appendix.

Step 5: Link variables

Copy the sentence text from the text box in the right sidebar and paste it into the search box of the recommendation tool. Select the appropriate document ID (which is the file name listed on the top of the opened document in INCEpTION) of the sentence and review the recommended variables. Note, that the recommendations are not 100% accurate and may not provide correct results in an accurate order. Link all relevant variables in the variable text box in the sidebar (e.g., "1a. Variable") by simply copying the variable ID (e.g., "ZA000_v1"). You can adjust the search input by using relevant parts of the sentence as input, or by thinking of re-phrased versions to find a suitable match.

If the suggestions from the variable recommender are insufficient, switch to the "Keyword search" page in the recommender. Here, you can input single or combinations of keywords (separated by a white-space). The suggestions will more closely resemble "exact-match" variables that contain the keywords.

Finally, go to the "Variable catalog" page to review all possible relevant variables. Note, however, that not all variables listed in the catalog are referenced in the publication, and that in rare cases, certain variables or research datasets are referenced in publications that are not listed in the catalog. In cases where a sentence in the publication references a variable not listed in the catalog, simply input an "UNK" variable tag (which stands for "unknown") in the variable text box. If multiple identical variables (i.e., variables that have been reused from one dataset to another) over multiple years (e.g., as is the case for a number of variables in the ALLBUS dataset), only select one of the variables.

Step 6: Rate confidence

For each annotation of a variable mention, also provide a rating for the confidence of the annotation on the provided scale (0=not very confident, 3=very confident). If you are uncertain about the presence of a specific variable, link the variable and provide a low confidence score (e.g., 0 or 1).

Step 7: Label type

For each annotation of a variable mention, please also select which type of mention it is: explicit or implicit. An explicit mention can be self-contained in the sentence and, while it may depend on previous context, but can be fully disambiguated (i.e., mapped to one or more variables in survey datasets). An implicit mention may require context, world or background knowledge, or be redefined with a different term (e.g., an indicator or concept variable, which may be made up of a set of variables) to understand. See examples for explicit and implicit variable mentions in the Appendix.

Step 8: Lock the document

Once you are done with a document, lock the document by clicking on the "Finish document" button (open lock icon). After this step, no more changes to the document annotation are possible. Continue annotating the next documents until there are no more documents to annotate.

Appendix

Examples

1. **Self-containing references within one single sentence**: Different references to the same variable occur in one paragraph. Both references are self-contained and valid.

> **Reference 1**: “To test this, we analyzed data on the strength of individuals’ identification with their home town and its inhabitants from the German ALLBUS surveys.
>
> **Reference 2**: “This is presented in figure II, which reports the marginal effect of being Catholic on the propensity to feel strongly attached to one’s home town and its inhabitants...’ ”
>
> **Variable label**: IDENTIFICATION WITH OWN COMMUNITY

2. **Context is necessary to determine the variable**: The sentences before/after the reference sentence may provide additional contextual information and valuable clues regarding the identification and disambiguation of variables.

> **Document Title**: “Exploring Sources of Punitiveness Among German Citizens
>
> **Section Title**: “Covariates of punitive attitudes.
>
> **Sentence before**: “For both, respondents were asked whether they agree or disagree with these statements, and the responses were recoded so that a positive response (1) indicates feelings of the designated type of cynicism.”
>
> **Reference**: “We also include a measure that we label as “life satisfaction”, which is a four- category item asking respondents the following: “All things considered, have your ideas of what you wanted to achieve in life been (1) more than fulfilled, (2) fulfilled, (3) not quite fulfilled, and (4) not at all fulfilled?”
>
> **Variable label**: PERSONAL AMBITIONS IN LIFE FULFILLED?

> **Document Title**: “RACISM IN SOCCER ELIMINATING SOCCER RACISM AND USING SPORT AS A VEHICLE FOR NATIONAL CHANGE”
>
> **Sentence before**: “This graph shows significantly lower levels of positive national pride and significantly higher levels of negative national pride in Germany compared to other comparable countries.”
>
> **Reference**: “There was an increase from 71% of the population stating they were “very proud” or “fairly proud” a few months before the games to 78% of the population stating they possessed these positive feelings during the games.”
>
> **Variable label**: v247 PROUD TO BE A GERMAN?

(a) **Linguistic variations - Quotation**: Variables may be directly quoted.

Reference: “There is only one item measuring happiness which directly asks the respondents: ‘If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole...’ ”

Variable question: “If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole...”

(b) **Linguistic variations - Paraphrase**: Variables may be paraphrased.

Reference: “The second and the third questions come from the ISSP research, where respondents were asked about **the influence** of religious leaders on **people’s votes** and the **government**.”

Variable question 1: “How much do you agree or disagree with each of the following: Religious leaders should not try to influence **how people vote** in elections.”

Variable question 2: “How much do you agree or disagree with each of the following: Religious leaders should not try **to influence government decisions**.”

(c) **Linguistic variation - Negative Polarity Item**: Variables may be expressed by means of a different polarity marker.

Reference: “Victimization and fear of crime are dichotomous, with “1” indicating positive responses to either of the two following questions: “Have you been a victim of theft in the past 3 years?” and “Is there any place in the immediate vicinity in which you **fear** walking alone at night?”

Variable question: “Is there any area in the immediate vicinity - I mean within a kilometer or so - where you would **prefer not to walk** alone at night?”

(d) **Linguistic variation - Lexical Inference**: Variables may require world or linguistic knowledge and awareness of the context.

Reference: “First, winners are more politically satisfied compared with losers, including those who voted for the **Black–Red Grand Coalition**.”

Variable question: “Which party did you vote for with your second vote (”Zweitstimme”)?
Variable answers:

- Respondent didn’t vote
- **The Christian Democratic/Christian Social Union CDU/CSU**
- **The Social Democratic Party SPD**

(e) **Metadata Reference**: Variables may be referenced by the different fields they are made up of, including: question, sub-question or item categories, label, response categories, and topic. The sub-question is often most informative, while labels are often hard to understand.

i. **Elliptical question and sub-question/item category**: Certain variables have a generic question (often followed by ellipses, i.e., three dots) and can only be distinguished from other variables at the level of sub-question/item category.

> **Reference**: “To measure anti-immigrant sentiments, a four-item scale was created ($\alpha = .72$) from responses to questions regarding citizens’ beliefs about immigration for four groups: asylum seekers, EU workers, non-EU workers, and ethnic Germans.”
>
> **Variable question**: “The following questions deal with the entry as immigrants of various groups of people into Germany. What is your opinion about this?
> **Variable subquestion**: **What about ethnic Germans from Eastern Europe?**

ii. **Various variable candidates at different levels of specificity**: v321 has the correct level of specificity in the example below.

> **Reference**: “This is presented in figure II, which reports the marginal effect of being Catholic on the propensity to feel strongly attached to one’s home town and its inhabitants. “
>
> **Variable ID**: v321
> **Variable label**: IDENTIFICATION WITH OWN COMMUNITY
> **Variable question**: Now we would like to know how strongly you identify with your own town (community) and its inhabitants. Please use the card for your answers
> **Variable Subquestion**: Do you identify emotionally with your town very strongly, pretty strongly, only weakly or not at all?
>
> **Variable ID**: v322
> **Variable label**: IDENTIFICATION WITH FEDERAL STATE
> **Variable question**: Now we would like to know how strongly you identify with your own town (community) and its inhabitants. Please use the card for your answers
> **Variable Subquestion**: Do you identify emotionally with your federal state very strongly, pretty strongly, only weakly or not at all?

iii. **Variable Response categories**: Variable response categories may necessary to identify a variable.

Reference: “Statistics on the degree of spirituality provide a clearer picture: 8.4% describe themselves as very spiritual, 29.7% as moderately spiritual, 32.8% as slightly spiritual, and 29.1% as not spiritual at all.”

Variable question: “What best describes you:“
Answer categories:

- I follow a religion and consider myself to be a spiritual person interested in the sacred or the supernatural.
- I follow a religion, but don’t consider myself to be a spiritual person interested in the sacred or the supernatural.
- I don’t follow a religion, but consider myself to be a spiritual person interested in the sacred or the supernatural.
- I don’t follow a religion and don’t consider myself to be a spiritual person interested in the sacred or the supernatural.

iv. **Variable Label**: The variable label may contain information to correctly identify a variable (e.g., ”Variable label: AEQUIVALENZEINKOMMEN OECD - NEU, KAT” or control variables such as age or gender).

Reference: “We run least squares regressions of attractiveness on anthropometric measures and several groups of control variables, including age, region, year, interviewer fixed effects, number of children, and health status.

Variable label: RESPONDENT: AGE
Variable question: -
Answer categories:
- Refused
- No answer

(f) **Explicit vs Implicit variable mentions**: Variables may be mentioned explicitly, which are usually self-contained mentions, or implicitly, which often require additional knowledge about an author’s definition of a variable, the context provided in the publication, or other external source of information or data. The table below illustrates a number of explicit and implicit cases:

Type	**Variable Mention**	**Variable/Explanation**
Explicit Self-contained	**Reference**: To test this, we analyzed data on the strength of individuals’ identification with their hometown and its inhabitants from the German ALLBUS surveys.	**Variable label**: IDENTIFICATION WITH OWN COMMUNITY
Explicit Self-contained	**Reference**: This is presented in figure II, which reports the marginal effect of being Catholic on the propensity to feel strongly attached to one’s home town and its inhabitants...	**Variable label**: IDENTIFICATION WITH OWN COMMUNITY

Explicit Context-dependent	**Context**: For both, respondents were asked whether they agree or disagree with these statements, and the responses were recoded so that a positive response (1) indicates feelings of the designated type of cynicism. **Reference**: We also include a measure that we label as “life satisfaction”, which is a four-category item asking respondents the following: “All things considered, have your ideas of what you wanted to achieve in life been (1) more than fulfilled, (2) fulfilled, (3) not quite fulfilled, and (4) not at all fulfilled?	**Variable label**: PERSONAL AMBITIONS IN LIFE FULFILLED?
Explicit Quotation	**Reference**: There is only one item measuring happiness which directly asks the respondents: “If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole. . . ”	**Variable question**: If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole...
Explicit Paraphrase	**Reference**: The second and the third questions come from the ISSP research, where respondents were asked about the influence of religious leaders on people’s votes and the government.	**Variable question 1**: How much do you agree or disagree with each of the following: Religious leaders should not try to influence how people vote in elections. **Variable question 2**: How much do you agree or disagree with each of the following: Religious leaders should not try to influence government decisions.
Explicit Negation in quotation	**References**: Victimization and fear of crime are dichotomous, with “1” indicating positive responses to either of the two following questions: “Have you been a victim of theft in the past 3 years?” and “Is there any place in the immediate vicinity in which you fear walking alone at night?”	**Variable question**: Is there any area in the immediate vicinity - I mean within a kilometer or so - where you would prefer not to walk alone at night?
Implicit Context-dependent	**Context**: This graph shows significantly lower levels of positive national pride and significantly higher levels of negative national pride in Germany compared to other comparable countries. **Reference**: There was an increase from 71% of the population stating they were “very proud” or “fairly proud” a few months before the games to 78% of the population stating they possessed these positive feelings during the games.”	**Variable label**: PROUD TO BE A GERMAN?
Implicit Lexical inference	**Reference**: First, winners are more politically satisfied compared with losers, including those who voted for the Black–Red Grand Coalition.	**Variable question**: Which party did you vote for with your second vote (”Zweitstimme”)? **Variable answers**: - The Christian Democratic / Christian Social Union CDU/CSU - The Social Democratic Party SPD

Implicit Context-dependent	**Reference**: Wie man sieht, sind die Vorurteile in Griechenland, Belgien und Westdeutschland am höchsten und in Italien, Luxemburg und Spanien am niedrigsten ausgeprägt.	**Variable question 1**: For each of the following opinions, please tell me whether you tend to agree or tend to disagree? **Item categories**: - People from these minority groups abuse the system of social benefits - The presence of people from these minority groups is a cause of insecurity - The presence of people from these minority groups increases unemployment in (COUNTRY) **Variable question 2**: Again, speaking generally about people from minority groups in terms of race, religion or culture, do you think there are not many, a lot but not too many, or too many of them living in (OUR COUNTRY)?
Implicit Part of concept or indicator variable(s)	**Reference**: Problematisch ist vor allem der Indikator ”Sozialsystem”.	Unknown set of variables making up the indicator ”Sozialsystem.”
Implicit Part of concept or indicator variable(s)	**Reference**: Für diese zehn Staaten wurde Modell 2 mit den Indikatoren Arbeitslosigkeit, Unsicherheit und Zahl geschätzt.	Unknown variables that make up the indicators ”Arbeitslosigkeit”, ”Unsicherheit”, and ”Zahl.”
Implicit - Context-dependent	**Reference**: Angegeben sind auch die 95%igen Konfidenzintervalle: die Vorurteile unterscheiden sich in einer ganzen Reihe von Staaten nicht signifikant.	Unknown set of variables that make up the indicator ”Vorurteile.”
Implicit - Part of concept or indicator variable(s)	**Reference**: Französische und irische Befragte stimmen dem Indikator Sozialsystem bei gleichen Einstellungen eher zu als Befragte anderer Staaten, weshalb die Zustimmung zu diesem Indikator nicht in gleichem Maße wie in anderen Staaten Vorurteile widerspiegelt.	Unknown set of variables making up the indicator ”Sozialsystem.”
Implicit - Context-dependent	**Reference**: Bemerkenswert ist zudem, dass nahezu die Hälfte aller griechischen und mehr als 40% der belgischen Befragten allen drei Aussagen zustimmt, wobei die Anteile der griechischen Befragten ohne bzw. mit geringen Vorurteilen vernachlässigt werden können.	Missing context for ”allen drei Aussagen” and unknown set of variables that make up the indicator ”Vorurteile.”
Implicit - Part of concept or indicator variable(s)	**Reference**: Eingangs wurden die Ergebnisse von Quillian (1995) berichtet, wonach ein erheblicher Teil der Unterschiede im Ausmaß von Vorurteilen zwischen Staaten (ca. 70%) durch die Größe der Minderheit und die makro-ökono-mische Lage erklärt werden.	Unknown set of variables that make up the indicator ”Vorurteile.”