Instructions for the data collection task
Non-finite complementation after the verb "help"

1. Find all instances of the root *help*

Your dataset must include:

- ALL instances of the root *help* (including nouns, adjectives like *helpless*, the verb *help* without an infinitive). We can then calculate how many instances of the root help are verbs, and how many of them occur with non-finite complements.
- the actual example as a KWIC concordance.
  ~120 characters before *help*, 240 characters after *help*
  ~15 raw words before *help*, 30 raw words after *help*
- "Hit" as the very first column. It should give every KWIC example a unique number.

Schematically, the beginning of your dataset could look like this:

Hit	KWIC	...
1	Example 1	...
2	Example 2	...
...	...	...

2. Label the dependent variable

- You dataset must include a column for the dependent variable:
  TO (*help to do*), (1)
  BARE (*help do*), (2)
  ING (*help* + –ing), (3)
  INING (*help* + *in* + - ing), (4)

(1) Practice helps you to get your timing down
(Brown, 1961, A-PressReport)

(2) he had helped fight an oil-well fire that raged six days and nights
(Brown, 1961, K-GeneralFiction)

(3) I could not help thinking of it
(COHA, 1870, StoryBadBoy)

(4) She was particularly pleased with being allowed to help in getting breakfast or tea
(COHA, 1850, MaryErskine)

- consider 15 words to the right of *help*
- Every example you extract must be coded for one of these four complementation patterns. If there is no complement clause or *help* is not a verb, label as 'NA'

Hit	KWIC	DepVar	...
1	Example 1	TO	...
2	Example 2	BARE	...
...	...	...	...

3. Extract example-level independent variables

You should value every example for the following independent variables:

- **Inflectional** form of *help*
  - This could be infinitive help, present tense help, 3rd person singular helps, past helped, past participle helped or present participle helping
  - The Penn tags for those are VB, VBP, VBZ, VBD, VBN, VBG
  - If you have a choice, use the most detailed POS-tag (e.g. CLAWS-5 rather than Universal tags in the BNC).
  - Use the Penn tagset if you POS-tag your corpus yourself
- Verb **lemma** of the non-finite complement clause (*help supply, help to find, help understand*, etc.). Consider 15 words to the right of help. Make sure you do not count punctuation as a word. If there is no non-finite complement, code this as 'NA.'
- The number of words **intervening** between *help* and the head of its non-finite complement. Make sure you do not count the non-finite marker *to* itself (or *in* in *help in doing something*).
  - Examples:
    - *She wanted to help her nice friend a lot to survive* = 5
    - *My parents help me understand the truth* = 1
    - *Money helps solve problems* = 0
    - *Money helps to solve problems* = 0
    - *Money helps in solving problems* = 0

  If there is no non-finite complement, code this as 'NA.'
- Four additional columns for the **object** of *help*:
  - (1) Is an object present or not? Code as (`Yes, No`). If there is non non-finite complement clause after *help*, code this as 'NA.'
  - (2) If an object is present, is its type either a pronoun or a full noun phrase? Code this as (`PRO, NP`). If there is no object, code this as 'NA.'
    - `PRO` includes all pronouns, *someone, who, myself, this*
  - (3) If an object is present, what is the length of the object (measured in words). Code as (`1, 2, 3, ...`). If there is no object, code this as 'NA.'
  - (4) If object present, what is the head of the object. Code this as the string (*them, children*, ...). If there is no non-finite complement, code this as 'NA.'
- **Voice** of *help*, passive (*to be helped,* be + VBN) vs. active ("normal" *help*). Consider 2 words to the left of *help*. Code as (Active, Passive). If *help* is not a verb, code this as 'NA'.
- **Preceding *to*** (horror aequi). Is there a non-finite marker *to* in front of *help*: *to help* or not? Consider 2 words to the left of *help* (e.g. *to never help*). Code this as (`NOtoBefore, YEStoBefore`). If *help* does not function as a verb, code this as 'NA.'
- **Polarity**: Is help negated or positive? For example, negative *help* could be *cannot help, does not help, can't help*, etc. Code this as (NEG, POS).

- Two additional columns for the **subject** of *help*. Look up to 30 words to the left of *help* to find its subject. Consider the subject in verb phrase coordination (*She came and helped me to survive*. – the subject is *she*):
    - (1) If *help* functions as a verb, what is the type of its subject? Code as one of the following types:
        - NP – full noun phrase (*children, Bill, my parents*)
            - This also includes gerunds / -ing forms (*Sleeping helps*)
            - For names, e.g. *Bill Smith*, use the family name
            - For coordination structures, use the first conjunct (e.g. use *Mary* in *Mary and Bill helped them explore the house.*)
        - It – the pronoun *it*
        - PRO – any other pronoun (*he, this, who*)
            - also in relative clauses (*the thing which helped me swim*)
        - NULL – there is no overt subject

          This can happen:
            - in a "diary" style, where subjects are left out (*Came home. Watched TV. Helped mum make dinner*).
            - in imperatives (*Help me!*)
            - with non-finite help in raising and control structures (*She promised me to help study*)
    - (2) If a subject type was added, what is its head? Code this as the string (*them, children*, ...). If this hit it not an instance of verbal *help*, code it as 'NA.'

4. Extract text-level meta-data

For every example, get the following text information:

COCA
textID
Text title
Year (2000-2012)
Genre (if the genre is SPOK, recode it as "TVInterview")
Mode ("Written" or "Spoken")
Variety (always "AmE")
Corpus (always "COCA")

COHA
TextID
TextTitle
Year (between 1800 and 1999)
Genre (MAG, NEWS, FIC, NF)
Mode (always "Written")
Variety (always "AmE")
Corpus (always "COHA")

BNC

TextID

Year of creation (typically around 1990, or 0000 if unknown)

Genre (e.g. NEWS)

Subgenre (e.g. newsp other: science)

Mode ("Written" or "Spoken")

Variety (Always "BrE")

Corpus (always BNC)

Spoken BNC 2014

TextID

SpeakerID

Year (always "2014")

SpeakerGender

Exact age or if not available, mid point of age range, (or "unknown")

Socgrade (A, B, C, D, E)

Variety (Always BrE)

Mode (Always "Spoken")

HUM19UK
TextTitle
Year
Author name
Author gender
Genre (always "Fiction")
Mode (always "Written")
Variety (Always "BrE")
Corpus (always "HUM19UK")

CLMET-3
TextID
Text name
Year of composition
Author name
Author gender
Genre (Drama, Letters, …)
Mode (always "Written")
Variety (Always "BrE")
Corpus (always “CLMET-3”)

TenIndivCorpus
Text title
Year of composition
Author name
Author gender
Genre (always "Fiction")
Mode (always "Written")
Variety (Always "BrE")
Corpus (always "TenIndivCorpus")

5. Some additional challenges

- Can you avoid counting punctuation as intervening words?
  Respect clause boundaries (punctuation).
  e.g. *What is the world coming to? Help me!*
- For ING and INING, try to change the -ing form (*deciding*) to the infinitive (*decide*) for the lemma of the head of the non-finite clause.
- For negation, can you rule out “not only”? … *not only help but also support* – POS
- Can you avoid non-finite clauses that depend on nouns, not verbs?
  e.g., … *to understand how these internal contradictions are reconciled, helps them in their [struggle to achieve personal salvation]* …

6. Instructions for your presentation

- Prepare slides.
- Bring along your script and discuss your code.
- 1-2 sentences on corpus content / material.
- Bibliographical reference for the corpus.
- Cite examples from your dataset.
- Mention how long it took your final script to run on the corpus.
- Arrange the final dataset in a spreadsheet. Format it professionally. Show it at the end of your presentation.
- A few remarks on evaluation. No formal evaluation is required (no accuracy, precision, F-score), but you must look at a fair number of examples and judge the quality of the labelling of your variables impressionistically.

The oral presentation will take place in about one month, on 19/3/2025.
However, you should start with this task starting now.

7. Advice

- Make sure you develop your script step-by-step
    - Start your script with one text.
    - Then move to a few text files.
    - Then move to a subsection (10% of corpus). Lots of testing!
    - When you're confident your script works, run it on the whole corpus.
- Become familiar with the structure of your corpus. Every project is different.
- See the dataset `HelpDatasetBrown.xlsx` for an illustration.
- It can be a really great feeling when the data collection works out. Empowering. Enjoy this task!