As part of your language investigation, you will need to collect data. In these activities, you will

- learn how to collect data for your language investigation
- consider possible advantages/disadvantages to different types of data
- consider other important issues when tackling your linguistic data.

Collecting data

Fill in the table below as thoroughly as you can. If you are aware of any disadvantages beforehand, you can consider this in your planning and take it into account when collecting your data.

Data types	Where/how would you collect this?	Advantages of this type of data?	Disadvantages of this type of data?
Spoken language (transcripts)			
Written/ multimodal language from print sources			
Written/ multimodal language from online sources			

Questions to consider before and during collecting your data

Make planning notes for your investigation, using the questions below as prompts. It's not essential for you to answer every question, but the more notes you make now, the clearer you'll be when it comes to doing your own project.

Technology	What technology do you need? Do you have access to it and do you know how to use it (e.g. voice recording software on a phone, concordancing software and other corpus tools)?
Quantity of data	How much data do you need? If you wish to investigate how a tabloid reports the same news topics compared to a broadsheet, how many articles should you collect? Is your data going to be short texts (e.g. tweets, advertising slogans)? You will need more texts than someone investigating political speeches. If your data consists of very short texts (e.g. advertising slogans) will there be enough to discuss all levels of the language levels, or are you going to focus on a particular language level (e.g. semantics/pragmatics)? If so, you will probably need more data than a student applying a whole language levels approach to their data.
Transcribing	How patient are you? How good are your IT and typing skills? Transcribing your own recordings can be time consuming, but the act of transcribing will make you very familiar with your data and can give you a head start on the next steps of an investigation.
Online transcripts	There are plenty of transcripts available online if you would like to investigate speech, but don't want to transcribe – are there already existing transcriptions available with regard to your chosen topic?
Combining data	You can combine different kinds of data – would your project benefit from a mixed approach? If you wish to investigate the language of a particular person or group, you may collect spoken data as well as written data produced by this person/group of people.

Ethical issues	Are there any ethical issues affecting the collection of your data? For example, if you wish to record people's speech, you need to have their permission. You can produce a simple permission slip for any participants to sign. This is especially important when transcribing children's speech – you must have the permission of a parent or guardian. Your teacher can help with this.
Access to data	If you wish to use existing data, such as a particular corpus, do you have access to it?
How interested are you?	Can you see yourself working on this topic/data for weeks? Will this project sustain your interest? You will naturally be more engaged with something you enjoy and are far more likely to produce an interesting and comprehensive investigation.
Why?	Lastly, and maybe most importantly, consider the 'so what?' question. Why are you studying the data? Why would anyone want to read your project? Why would they be interested in your results?

Teaching notes

Every method of data collection has its advantages and disadvantages – it is beneficial for students to anticipate the disadvantages of their chosen data type(s) so that they can find ways of working around these.

Spoken data is often popular – because it can yield so much for analysis – but for some students the act of transcription is too daunting. However, there are resources available where students can access spoken data already transcribed:

- Hansard: although this is 'cleaned up', it is still interesting material with regard to language and power, the language of politics for example.

- Talkbank is a database where transcripts can be downloaded: talkbank.org/ and incorporates the Child Language Acquisition database childes.talkbank.org/ – it also has multilingual/language learner data.

- Spoken language corpora are available online often through universities – the University of California, Santa Barbara makes its corpus of spoken American English freely available via its website – transcripts are organised according to topic and can be downloaded as text files. www.linguistics.ucsb.edu/research/santa-barbara-corpus.