What is the main goal of collecting real language data?

The main goal of collecting real language data is to construct accurate descriptions of the data and to make linguistic generalisations about the data. In terms of syntax, this means collecting examples of both acceptable and unacceptable data, not just acceptable data.

How do I know what to ask?

Initially, the theory should be your guide as to what to ask. There are some major thematic areas in the theory which can drive a lot of the basic description. Each basic area of the theory has some kinds of data which have driven it. Sometimes this data is in English and sometimes it is in other languages. As a starting point, it makes sense to try to get versions of this ‘core’ data of an area in the language you are studying. Sometimes there will not be an equivalent sentence: this then leads to more questions. As soon as you have a small piece of data in the language, however, you should always try to ask for more data using the examples you have already found instead of asking for translations. For example, suppose you have elicited example (1a). If you suspect that word order might be variable, you might want to ask (1b) or (1c) simply by changing the order yourself and then asking the question. If you just asked “How do you say ‘who gave what to whom?’” you might simply elicit (1a) and miss (1b/c).

(1) a. Ki kinek mit adott? (Hungarian)
   who to whom what gave
   Who gave what to whom?
   b. Ki mit kinek adott?
   who what to whom gave
   Who gave what to whom?
   c. Kinek mit ki adott?
   to whom what who gave
   Who gave what to whom?

Remember, if you can say something in one language, there should be a way to express the same thought in another language, and you should try to push those intuitions with your consultants. As you collect data, you should be constantly constructing hypotheses to use to ask the next question. The notion of a ‘minimal pair’, is also fundamental. In syntax a minimal pair is two sentences that differ by one syntactic property. For example, (2a/b) form a minimal pair, as do (3a/b), but (2a/3a) are also a minimal pair, and (2b/3b).

(2) a. John hoped to leave.
   b. *John hoped Bill to leave.

(3) a. John wanted to leave.
   b. John wanted Bill to leave.
How much linguistics does my consultant need to know?

If your consultant knows no linguistics at all, you need to teach enough about the kinds of judgements you are looking for, etc. so that the speaker doesn’t give you erroneous data based on some misconceptions about the questions you are asking.

How do I present my data?

First of all, the data should be organised in some understandable way: data that is similar should be grouped together, minimal pairs should be presented together etc. If the data is being organised in terms of an analysis, the order of the data should follow the order of the argument made to defend the analysis. On a more detailed note, non-English data needs to be represented in three lines of text as in the Hungarian example above and the following example.

(4) Os homen-s sa´ıram da sala. (Portuguese)
    the.M.PL man-PL left.3PL from-the.F.SG room.

    The men left the room.

1. The first line presents a transliterated or written version of the sentence. If the language is morphologically complex, then it is usually normal to break apart the individual morphemes in the first line, separated by hyphens.

2. The second line (called the ‘gloss’) presents a word by word or morpheme by morpheme translation of the sentence. You should make sure that the glosses line up with the words/morphemes they are glossing. You should use tabs and not spaces to do this, as in the example above. Glosses should include relevant morphological analysis of complex forms. There are a couple of conventions to this. If a single morpheme encodes more than one feature, the separate features are usually connected together with a period (‘.’) This is shown in (4) with the morpheme ‘os’ which is glossed as ‘the.M.PL’. Since none of the features are separate morphemes from ‘the’, the features are added without hyphens. In the case of the word ‘homens’ the plural morpheme is a separate morpheme that is glossed with a hyphen: ‘homen-s’ glossed as ‘man-PL’. If a particular element has no obvious translation, it is best to simply gloss it as itself. For example, Chinese has a morpheme BA that shows up in front of a direct object that has moved to preverbal position, as in (5). Since it is not clear what BA actually means, it is safer to just gloss it as BA rather than invent a meaning which is likely to be wrong. Chinese also has a morpheme ‘le’ which encodes aspect. This is glossed in the example as ‘ASP’ rather than giving it a determinate meaning.

(5) Ta ba gonke zuowan-le (Chinese)
    he BA homework do-finish-ASP
    He finished his homework

3. The third line gives an idiomatic English translation of the sentence. As you can see in (3) and (4) this translation does not necessarily match the word for word translation exactly. The amount of detail needed in a gloss varies somewhat depending on what you are describing, since it gets tedious to gloss every agreement feature on every phrase of every example. If you decide not to gloss each morpheme, you need to provide some table of the relevant paradigm. This is especially true for agreement morphology in languages that have agreement.
Don’t exploit your consultant

Your consultant’s job is to answer your questions, not to tell you everything they know about their language. This is especially true for those of you who are linguists: although you can be helpful, you should not think of your task as a consultant to teach about your language, but simply to answer questions and supply data.

Some pitfalls

Remember that judgements depend on context, meaning, pragmatics, processing and a whole bunch of factors that are not directly connected to ‘grammaticality’. Also, speakers (especially non-linguists) are often willing to suggest ‘alternatives’ to a sentence you ask. These alternatives, may not really answer the question you are asking. For example, if you ask a native English speaker “Can you say ‘John, I like’” they might say that it would be ‘better’ to say “I like John”. But this isn’t the question you asked. And in fact you can say “John, I like” but you need a particular context to do it. Another pitfall that non-linguist consultants sometimes create is that they want to be helpful, so they say that most things are ok, even when they are not. Some of the first data collected by Lisa Travis, a linguist who works on Malagasy (which is a VOS language) when she was a student, turned out to be ‘foreigner talk’, i.e. sentences that were SVO, just to keep her happy! So be careful!