Corpus linguistics

A guide to the methodology

Anatol Stefanowitsch

Textbooks in Language Sciences 7

language science press

1 The need for corpus data

Broadly speaking, science is the study of some aspect of the (physical, natural or social) world by means of systematic observation and experimentation, and linguistics is the scientific study of those aspects of the world that we summarize under the label *language*. Again very broadly, these encompass, first, language *systems* (sets of linguistic elements and rules for combining them) as well as mental representations of these systems, and second, *expressions* of these systems (spoken and written utterances) as well as mental and motorsensory processes involved in the production and perception of these expressions. Some linguists study only the linguistic system, others study only linguistic expressions. Some linguists study linguistic systems as formal entities, others study them as mental representations. Some linguists study linguistic expressions in their social and/or cultural contexts, others study them in the context of production and comprehension processes. Everyone should agree that whatever aspect of language we study and from whatever perspective we do so, if we are doing so scientifically, observation and experimentation should have a role to play.

Let us define a corpus somewhat crudely as a large collection of authentic text (i.e., samples of language produced in genuine communicative situations), and corpus linguistics as any form of linguistic inquiry based on data derived from such a corpus. We will refine these definitions in the next chapter to a point where they can serve as the foundation for a methodological framework, but they will suffice for now.

Defined in this way, corpora clearly constitute recorded observations of language behavior, so their place in linguistic research seems so obvious that anyone unfamiliar with the last sixty years of mainstream linguistic theorizing will wonder why their use would have to be justified at all. I cannot think of any other scientific discipline whose textbook authors would feel compelled to begin their exposition by defending the use of observational data, and yet corpus linguistics textbooks often do exactly that.

The reasons for this defensive stance can be found in the history of the field, which until relatively recently has been dominated by researchers interested mainly in language as a formal system and/or a mental representation of such

a system. Among these researchers, the role of corpus data, and the observation of linguistic behavior more generally is highly controversial. While there are formalists who have discovered (or are beginning to discover) the potential of corpus data for their research, much of the formalist literature has been, and continues to be, at best dismissive of corpus data, at worst openly hostile. Corpus data are attacked as being inherently flawed in ways and to an extent that leaves them with no conceivable use at all in linguistic inquiry.

In this literature, the method proposed instead is that of *intuiting* linguistic data. Put simply, intuiting data means inventing sentences exemplifying the phenomenon under investigation and then judging their *grammaticality* (roughly, whether the sentence is a possible sentence of the language in question). To put it mildly, inventing one's own data is a rather subjective procedure, so, again, anyone unfamiliar with the last sixty years of linguistic theorizing might wonder why such a procedure was proposed in the first place and why anyone would consider it superior to the use of corpus data.

Readers familiar with this discussion or readers already convinced of the need for corpus data may skip this chapter, as it will not be referenced extensively in the remainder of this book. For all others, a discussion of both issues – the alleged uselessness of corpus data and the alleged superiority of intuited data – seems indispensable, if only to put them to rest in order to concentrate, throughout the rest of this book, on the vast potential of corpus linguistics and the exciting avenues of research that it opens up.

Section 1.1 will discuss four major points of criticisms leveled at corpus data. As arguments against corpus data, they are easily defused, but they do point to aspects of corpora and corpus linguistic methods that must be kept in mind when designing linguistic research projects. Section 1.2 will discuss intuited data in more detail and show that it does not solve any of the problems associated (rightly or wrongly) with corpus data. Instead, as Section 1.3 will show, intuited data actually creates a number of additional problems. Still, intuitions we have about our native language (or other languages we speak well) can nevertheless be useful in linguistic research – as long as we do not confuse them with "data".

1.1 Arguments against corpus data

The four major points of criticism leveled at the use of corpus data in linguistic research are the following:

1. corpora are usage data and thus of no use in studying linguistic knowledge;
2. corpora and the data derived from them are necessarily incomplete;

3. corpora contain only linguistic forms (represented as graphemic strings), but no information about the semantics, pragmatics, etc. of these forms; and

4. corpora do not contain negative evidence, i.e., they can only tell us what is possible in a given language, but not what is not possible.

I will discuss the first three points in the remainder of this section. A fruitful discussion of the fourth point requires a basic understanding of statistics, which will be provided in Chapters 5 and 6, so I will postpone it and come back to it in Chapter 8.

1.1.1 Corpus data as usage data

The first point of criticism is the most fundamental one: if corpus data cannot tell us anything about our object of study, there is no reason to use them at all. It is no coincidence that this argument is typically made by proponents of generative syntactic theories, who place much importance on the distinction between what they call *performance* (roughly, the production and perception of linguistic expressions) and *competence* (roughly, the mental representation of the linguistic system). Noam Chomsky, one of the first proponents of generative linguistics, argued early on that the exclusive goal of linguistics should be to model competence, and that, therefore, corpora have no place in serious linguistic analysis:

> The speaker has represented in his brain a grammar that gives an ideal account of the structure of the sentences of his language, but, when actually faced with the task of speaking or "understanding", many other factors act upon his underlying linguistic competence to produce actual performance. He may be confused or have several things in mind, change his plans in midstream, etc. Since this is obviously the condition of most actual linguistic performance, *a direct record – an actual corpus – is almost useless, as it stands, for linguistic analysis of any but the most superficial kind* (Chomsky 1964: 36, emphasis mine).

This argument may seem plausible at first glance, but it is based on at least one of two assumptions that do not hold up to closer scrutiny: first, that there is an impenetrable bi-directional barrier between competence and performance, and second, that the influence of confounding factors on linguistic performance cannot be identified in the data.

The assumption of a barrier between competence and performance is a central axiom in generative linguistics, which famously assumes that language acquisition depends on input only minimally, with an innate "universal grammar" doing

most of the work. This assumption has been called into question by a wealth of recent research on language acquisition (see Tomasello 2003 for an overview). But even if we accept the claim that linguistic competence is not derived from linguistic usage, it would seem implausible to accept the converse claim that linguistic usage does not reflect linguistic competence (if it did not, this would raise the question what we need linguistic competence for at all).

This is where the second assumption comes into play. If we believe that linguistic competence is at least broadly reflected in linguistic performance, as I assume any but the most hardcore generativist theoreticians do, then it should be possible to model linguistic knowledge based on observations of language use – unless there are unidentifiable confounding factors distorting performance, making it impossible to determine which aspects of performance are reflections of competence and which are not. Obviously, confounding factors exist – the confusion and the plan-changes that Chomsky mentions, but also others like tiredness, drunkenness and all the other external influences that potentially interfere with speech production. However, there is no reason to believe that these factors and their distorting influence cannot be identified and taken into account when drawing conclusions from linguistic corpora.[1]

Corpus linguistics is in the same situation as any other empirical science with respect to the task of deducing underlying principles from specific manifestations influenced by other factors. For example, Chomsky has repeatedly likened linguistics to physics, but physicists searching for gravitational waves do not reject the idea of observational data on the basis of the argument that there are "many other factors acting upon fluctuations in gravity" and that therefore "a direct record of such fluctuations is almost useless". Instead, they attempt to identify these factors and subtract them from their measurements.

In any case, the gap between linguistic usage and linguistic knowledge would be an argument against corpus data only if there were a way of accessing linguistic knowledge directly and without the interference of other factors. Sometimes, intuited data is claimed to fit this description, but as I will discuss in Section 1.2.1, not even Chomsky himself subscribes to this position.

[1] In fact, there is an entire strand of experimental and corpus-based research that not only takes disfluencies, hesitation, repairs and similar phenomena into account, but actually treats them as object of study in their own right. The body of literature produced by this research is so large that it makes little sense to even begin citing it in detail here, but cf. Kjellmer (2003), Corley & Stewart (2008) and Gilquin & De Cock (2011) for corpus-based approaches.

1.1.2 The incompleteness of corpora

Next, let us look at the argument that corpora are necessarily incomplete, also a long-standing argument in Chomskyan linguistics:

> [I]t is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances obtained by the linguist in field work. Any grammar of a language will project the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances (Chomsky 1957: 15).

Let us set aside for now the problems associated with the idea of grammaticality and simply replace the word *grammatical* with *conventionally occurring* (an equation that Chomsky explicitly rejects). Even the resulting, somewhat weaker statement is quite clearly true, and will remain true no matter how large a corpus we are dealing with. Corpora are incomplete in at least two ways.

First, corpora – no matter how large – are obviously finite, and thus they can never contain examples of every linguistic phenomenon. As an example, consider the construction [*it doesn't matter the* N] (as in the lines *It doesn't matter the colour of the car / But what goes on beneath the bonnet* from the Billy Bragg song *A Lover Sings*).[2] There is ample evidence that this is a construction of British English. First, Bragg, a speaker of British English, uses it in a song; second, most native speakers of English will readily provide examples if asked; third, as the examples in (1) show, a simple web query for ⟨ `"it doesn't matter the"` ⟩ will retrieve hits that have clearly been produced by native speakers of British English and other varieties (note that I enclose corpus queries in angled brackets in order to distinguish them from the linguistic expressions that they are meant to retrieve from the corpus):

(1) a. *It doesn't matter the reasons* people go and see a film as long as they go and see it. (thenorthernecho.co.uk)

b. Remember, *it doesn't matter the size of your garden*, or if you live in a flat, there are still lots of small changes you can make that will benefit wildlife. (avonwildlifetrust.org.uk)

[2]Note that this really is a grammatical construction in its own right, i.e., it is not a case of right-dislocation (as in *It doesn't matter, the color* or *It is not important, the color*). In cases of right-dislocation, the pronoun and the dislocated noun phrase are co-referential and there is an intonation break before the NP (in standard English orthographies, there is a comma before the NP). In the construction in question, the pronoun and the NP are not co-referential (*it* functions as a dummy subject) and there is no intonation break (cf. Michaelis & Lambrecht 1996 for a detailed (non-corpus-based) analysis of the very similar [*it BE amazing the* N]).

c. *It doesn't matter the context.* In the end, trust is about the person extending it. (clocurto.us)
d. *It doesn't matter the color of the uniform*, we all work for the greater good. (fw.ky.gov)

However, the largest currently publicly available linguistic corpus of British English, the one-hundred-million-word British National Corpus, does not contain a single instance of this construction. This is unlikely to be due to the fact that the construction is limited to an informal style, as the BNC contains a reasonable amount of informal language. Instead, it seems more likely that the construction is simply too infrequent to occur in a sample of one hundred million words of text. Thus, someone studying the construction might wrongly conclude that it does not exist in British English on the basis of the BNC.

Second, linguistic usage is not homogeneous but varies across situations (think of the kind of variation referred to by terms such as *dialect*, *sociolect*, *genre*, *register*, *style* etc., which I will discuss in more detail in Section 2.1 below). Clearly, it is, for all intents and purposes, impossible to include this variation in its entirety in a given corpus. This is a problem not only for studies that are interested in linguistic variation but also for studies in core areas such as lexis and grammar: many linguistic patterns are limited to certain varieties, and a corpus that does not contain a particular language variety cannot contain examples of a pattern limited to that variety. For example, the verb *croak* in the sense 'die' is usually used intransitively, but there is one variety in which it also occurs transitively. Consider the following representative examples:

(2) a. Because he was a skunk and a stool pigeon ... I *croaked* him just as he was goin' to call the bulls with a police whistle ... (Veiller, *Within the Law*)
b. [Use] your bean. If I had *croaked the guy* and frisked his wallet, would I have left my signature all over it? (Stout, *Some Buried Cesar*)
c. I recall pointing to the loaded double-barreled shotgun on my wall and replying, with a smile, that I would *croak at least two of them* before they got away. (Thompson, *Hell's Angels*)

Very roughly, we might characterize this variety as *tough guy talk*, or perhaps *tough guy talk as portrayed in crime fiction* (I have never come across an example outside of this (sub-)genre). Neither of these varieties is prominent among the

text categories represented in the BNC, and therefore the transitive use of *croak* 'die' does not occur in this corpus.[3]

The incompleteness of linguistic corpora must therefore be accepted and kept in mind when designing and using such a corpus (something I will discuss in detail in the next chapter). However, it is not an argument against the use of corpora, since *any* collection of data is necessarily incomplete. One important aspect of scientific work is to build general models from incomplete data and refine them as more data becomes available. The incompleteness of observational data is not seen as an argument against its use in other disciplines, and the argument gained currency in linguistics only because it was largely accepted that intuited data are more complete. I will argue in Section 1.2.2, however, that this is not the case.

1.1.3 The absence of meaning in corpora

Finally, let us turn to the argument that corpora do not contain information about the semantics, pragmatics, etc. of the linguistic expressions they contain. Lest anyone get the impression that it is only Chomskyan linguists who reject corpus data, consider the following statement of this argument by George Lakoff, an avowed anti-Chomskyan:

> Corpus linguistics can only provide you with utterances (or written letter sequences or character sequences or sign assemblages). To do cognitive linguistics with corpus data, you need to interpret the data – to give it meaning. The meaning doesn't occur in the corpus data. Thus, introspection is always used in any cognitive analysis of language [...] (Lakoff 2004).

Lakoff (and others putting forward this argument) are certainly right: if the corpus itself was all we had, corpus linguistics would be reduced to the detection of formal patterns (such as recurring combinations) in otherwise meaningless strings of symbols.

There are cases where this is the best we can do, namely, when dealing with documents in an unknown or unidentifiable language. An example is the *Phaistos disc*, a clay disk discovered in 1908 in Crete. The disc contains a series of symbols that appear to be pictographs (but may, of course, have purely phonological value), arranged in an inward spiral. These pictographs may or may not

[3] A kind of pseudo-transitive use with a dummy object does occur, however: *He croaked it* meaning 'he died', and of course the major use of *croak* ('to speak with a creaky voice') occurs transitively.

present a writing system, and no one knows what language, if any, they may represent (in fact, it is not even clear whether the disc is genuine or a fake). However, this has not stopped a number of scholars from linguistics and related fields from identifying a number of intriguing patterns in the series of pictographs and some general parallels to known writing systems (see Robinson (2002: ch. 11) for a fairly in-depth popular account). Some of the results of this research are suggestive and may one day enable us to identify the underlying language and even decipher the message, but until someone does so, there is no way of knowing if the theories are even on the right track.

It hardly seems desirable to put ourselves in the position of a Phaistos disc scholar artificially, by excluding from our research designs our knowledge of English (or whatever other language our corpus contains); it is quite obvious that we should, as Lakoff (2004) says, interpret the data in the course of our analysis. But does this mean that we are using introspection in the same way as someone inventing sentences and judging their grammaticality?

I think not. We need to distinguish two different kinds of introspection: (i) *intuiting*, i.e. practice of introspectively accessing one's linguistic experience in order to create sentences and assign grammaticality judgments to them; and (ii) *interpreting*, i.e. the practice of assigning an interpretation (in semantic and pragmatic terms) to an utterance. These are two very different activities, and there is good reason to believe that speakers are better at the second activity than at the first: interpreting linguistic utterances is a natural activity – speakers must interpret everything they hear or read in order to understand it; inventing sentences and judging their grammaticality is *not* a natural activity – speakers never do it outside of papers on grammatical theory. Thus, one can take the position that interpretation has a place in linguistic research but intuition does not. Nevertheless, interpretation is a subjective activity and there are strict procedures that must be followed when including its results in a research design. This issue will be discussed in more detail in Chapter 4.

As with the two points of criticism discussed in the preceding subsections, the problem of interpretation would be an argument against the use of corpus data only if there were a method that avoids interpretation completely or that at least allows for interpretation to be made objective.

1.2 Intuition

Intuited data would not be the only alternative to corpus data, but it is the one proposed and used by critics of the latter, so let us look more closely at this

practice. Given the importance of grammaticality judgments, one might expect them to have been studied extensively to determine exactly what it is that people are doing when they are making such judgments. Surprisingly, this is not the case, and the few studies that do exist are hardly ever acknowledged as potentially problematic by those linguists that routinely rely on them, let alone discussed with respect to their place in scientific methodology.

One of the few explicit discussions is found in Jackendoff (1994). Jackendoff introduces the practice of intuiting grammaticality judgments as follows:

> [A]mong the kinds of experiments that can be done on language, one kind is very simple, reliable, and cheap: *simply present native speakers of a language with a sentence or phrase, and ask them to judge whether or not it is grammatical in their language or whether it can have some particular meaning.* [...] The idea is that although we can't observe the mental grammar of English itself, we *can* observe the judgments of grammaticality and meaning that are produced by using it (Jackendoff 1994: 47, emphasis mine).

This statement is representative of the general assumptions underlying the practice of grammaticality judgments in generative linguistics (and many other frameworks) in two ways: first, in that it presents individual grammaticality judgments as a kind of scientific experiment on a par with more sophisticated experiments, and second, in that it presents grammaticality judgments as a direct reflection of a speaker's mental representation of the language in question.

Jackendoff briefly touches upon a crucial problem of the first assumption:

> Ideally, we might want to check these experiments out by asking large numbers of people under controlled circumstances, and so forth. But in fact the method is so reliable that, for a very good first approximation, linguists tend to trust their own judgments and those of their colleagues (Jackendoff 1994: 48).

It is certainly true that linguists trust their own judgments, but that does not mean, of course, that this trust is justified. There is little evidence that individual grammaticality judgments are reliable: in the linguistic literature, grammaticality judgments of the same sentences by different authors often differ considerably and the few studies that have investigated the reliability of grammaticality judgments have consistently shown that such judgments display too much variation within and across speakers to use them as linguistic data (cf., e.g., Schütze (1996) (reissued under a Creative-Commons license by Language Science Press in 2016), esp. Ch. 3 on factors influencing grammaticality judgments, and Cowart (1997)).

The attraction of grammaticality judgments lies not so much in their reliability, then, but in the ease with which they can be collected, and Jackendoff is very explicit about this when he says that

> other kinds of experiments can be used to explore properties of the mental grammar [...] Their disadvantage is their relative inefficiency: it takes a great deal of time to set up the experiment. By contrast, when the experiment consists of making judgments of grammaticality, there is nothing simpler than devising and judging some more sentences (Jackendoff 1994: 49).

However, the fact that something can be done quickly and effortlessly does not make it a good scientific method. If one is serious about using grammaticality judgments – and there are research questions that are not easily addressed without them –, then these judgments must be made as reliable as possible; among other things, this involves the two aspects mentioned by Jackendoff in passing: first, asking large numbers of speakers (or at least more than one) and, second, controlling the circumstances under which they are asked (cf. Schütze 1996 and Cowart 1997 for detailed suggestions as to how this is to be done and Bender 2005 for an interesting alternative; cf. also Section 4.2.3 in Chapter 4). In order to distinguish such empirically collected introspective data from data intuited by the researcher, I will refer to the former as *elicitation data* and continue to reserve for the latter the term *intuition* or *intuited "data"*.

In sum, there are serious problems with the reliability of linguistic intuition in general, a point I will briefly return to in Section 1.3. In the case of isolated judgments *by the researchers themselves*, these problems are compounded by two additional ones: first, the researchers are language experts, whose judgments will hardly be representative of the average native speaker – as Ronald Langacker has quipped (in an example sentence meant to illustrate syntactic complexity): "Linguists are no different from any other people who spend nineteen hours a day pondering the complexity of grammar [...]" (Langacker 1973: 109). Second, they will usually know what it is that they want to prove, and this will distort their judgments. Thus, expert judgments should be used with extreme caution (cf. Labov 1996) if at all (Schütze 1996), instead of serving as the default methodology in linguistics.

Let us return to the second assumption in the passage quoted above – that grammaticality judgments are transparently related to the mental grammar of the speaker producing them. In particular, let us discuss whether intuited "data" fare better than corpus data in terms of the three major points of criticism discussed in the preceding section:

1. Are intuited "data" a more direct reflection of linguistic knowledge (competence) than corpus data;
2. are intuited "data" more complete than corpus data; and
3. do intuited "data" contain information about the semantics, pragmatics, etc. of these forms.

1.2.1 Intuition as performance

The most fundamental point of criticism leveled against corpus data concerns the claim that since corpora are samples of language use ("performance"), they are useless in the study of linguistic knowledge ("competence"). I argued in Section 1.1.1 above that this claim makes sense only in the context of rather implausible assumptions concerning linguistic knowledge and linguistic usage, but even if we accept these assumptions, the question remains whether intuited judgments are different from corpus data in this respect.

It seems obvious that both inventing sentences and judging their grammaticality are kinds of behavior and, as such, performance in the generative linguistics sense. In fact, Chomsky himself admits this:

> [W]hen we study competence – the speaker-hearer's knowledge of his language – we may make use of his reports and his behavior as evidence, but we must be careful not to confuse "evidence" with the abstract constructs that we develop on the basis of evidence and try to justify in terms of evidence. [...] Since *performance – in particular, judgments about sentences – obviously involves many factors apart from competence*, one cannot accept as an absolute principle that the speaker's judgments will give an accurate account of his knowledge. (Chomsky 1972: 187, emphasis mine).

There is little to add to this statement, other than to emphasize that if it is possible to construct a model of linguistic competence on the basis of intuited judgments that involve factors other than competence, it should also be possible to do so on the basis of corpus data that involve factors other than competence, and the competence/performance argument against corpus data collapses.

1.2.2 The incompleteness of intuition

Next, let us turn to the issue of incompleteness. As discussed in Section 1.1.2, corpus data are necessarily incomplete, both in a quantitative sense (since every

corpus is finite in size) and in a qualitative sense (since even the most carefully constructed corpus is skewed with respect to the language varieties it contains). This incompleteness is not an argument against using corpora as such, but it might be an argument in favor of intuited judgments if there was reason to believe that they are more complete.

To my knowledge, this issue has never been empirically addressed, and it would be difficult to do so, since there is no complete data set against which intuited judgments could be compared. However, it seems implausible to assume that such judgments are more complete than corpus data. First, just like a corpus, the linguistic experience of a speaker is finite and any mental generalizations based on this experience will be partial in the same way that generalizations based on corpus data must be partial (although it must be admitted that the linguistic experience a native speaker gathers over a lifetime exceeds even a large corpus like the BNC in terms of quantity). Second, just like a corpus, a speaker's linguistic experience is limited to certain language varieties: most English speakers have never been to confession or planned an illegal activity, for example, which means they will lack knowledge of certain linguistic structures typical of these situations.

To exemplify this point, consider that many speakers of English are unaware of the fact that there is a use of the verb *bring* that has the valency pattern (or subcategorization frame) [*bring* $\text{NP}_{\text{LIQUID}}$ [$_{\text{PP}}$ *to the boil*]] (in British English) or [*bring* $\text{NP}_{\text{LIQUID}}$ [$_{\text{PP}}$ *to a boil*]] (in American English). This use is essentially limited to a single genre, – recipes: of the 145 matches in the BNC, 142 occur in recipes and the remaining three in narrative descriptions of someone following a recipe. Thus, a native speaker of English who never reads cookbooks or cooking-related journals and websites and never watches cooking shows on television can go through their whole life without encountering the verb *bring* used in this way. When describing the grammatical behavior of the verb *bring* based on their intuition, this use would not occur to them, and if they were asked to judge the grammaticality of a sentence like *Half-fill a large pan with water and bring to the boil* [BNC A7D], they would judge it ungrammatical. Thus, this valency pattern would be absent from their description in the same way that transitive *croak* 'die' or [*it doesn't matter the* N] would be absent from a grammatical description based on the BNC (where, as we saw in Section 1.1.2, these patterns do not occur).

If this example seems too speculative, consider Culicover's analysis of the phrase *no matter* (Culicover 1999: 106f.). Culicover is an excellent linguist by any standard, but he bases his intricate argument concerning the unpredictable nature of the phrase *no matter* on the claim that the construction [*it doesn't matter the* N] is ungrammatical. If he had consulted the BNC, he might be excused for

coming to this wrong conclusion, but he reaches it without consulting a corpus at all, based solely on his native-speaker intuition.[4]

1.2.3 Intuitions about form and meaning

Finally, let us turn to the question whether intuited "data" contain information about meaning. At first glance, the answer to this question would appear to be an obvious "yes": if I make up a sentence, of course I know what that sentence means. However, a closer look shows that matters are more complex and the answer is less obvious.

Constructing a sentence and interpreting a sentence are two separate activities. As a consequence, I do *not* actually know what my constructed sentence means, but only what I *think* it means. While I may rightly consider myself the final authority on the *intended* meaning of a sentence that I myself have produced, my interpretation ceases to be privileged in this way once the issue is no longer my intention, but the interpretation that my constructed sentence would conventionally receive in a particular speech community. In other words, the interpretation of a constructed sentence is subjective in the same way that the interpretation of a sentence found in a corpus is subjective. In fact, interpreting other people's utterances, as we must do in corpus linguistic research, may actually lead to more intersubjectively stable results, as interpreting other people's utterances is a more natural activity than interpreting our own: the former is what we routinely engage in in communicative situations, the latter, while not exactly unnatural, is a rather exceptional activity.

On the other hand, it is very difficult *not* to interpret a sentence, but that is exactly what I would have to do in intuiting grammaticality judgments – judging a sentence to be grammatical or ungrammatical is supposed to be a judgment purely about form, dependent on meaning only insofar as that meaning is relevant to the grammatical structure. Consider the examples in (3):

(3) a. When she'd first moved in she hadn't cared about anything, certainly not her surroundings – they had been the least of her problems – and *if the villagers hadn't so kindly donated her furnishings* she'd probably still be existing in empty rooms. (BNC H9V)

[4] Culicover is a speaker of American English, so if he were writing his book today, he might check the 450-Million-word Corpus of Contemporary American English (COCA), first released in 2008, instead of the BNC. If he did, he would find a dozen or more instances of the construction, depending which version he were to use – for example *It doesn't matter the number of zeros they attach to it*, from a 1997 transcript of ABC Nightline –, so he would not have to rely on his incomplete native-speaker intuition.

b. [$_{VP}$ *donated* [$_{NP}$ *her*] [$_{NP}$ *furnishings*]]
c. [$_{VP}$ *donated* [$_{NP}$ [$_{DET}$ *her*] [$_{N}$ *furnishings*]]]
d. Please have a look at our wish-list and see if you can *donate us a plant* we need. (headway-cambs.org.uk)

The grammaticality of the clause *[T]he villagers [...] donated her furnishings* in (3a) can be judged for its grammaticality only after disambiguating between the meanings associated with the structures in (3b) and (3c).

The structure in (3b) is a ditransitive, which is widely agreed to be impossible with *donate* (but see Stefanowitsch 2007a), so the sentence would be judged ungrammatical under this reading by the vast majority of English speakers. The structure in (3c), in contrast, is a simple transitive, which is one of the two most frequent valency patterns for *donate*, so the sentence would be judged grammatical by all English speakers. The same would obviously be true if the sentence was constructed rather than taken from a corpus.

But the semantic considerations that increase or decrease our willingness to judge an utterance as grammatical are frequently more subtle than the difference between the readings in (3b) and (3c).

Consider the example in (3d), which contains a clear example of *donate* with the supposedly ungrammatical ditransitive valency pattern. Since this is an authentic example, we cannot simply declare it ungrammatical; instead, we must look for properties that distinguish this example from more typical uses of *donate* and try to arrive an an explanation for such exceptional, but possible uses. In Stefanowitsch (2007a), looking at a number of such exceptional uses, I suggest that they may be made possible by the highly untypical sense in which the verb *donate* is used here. In (3d) and other ditransitive uses, *donate* refers to a direct transfer of something relatively valueless from one individual to another in a situation of personal contact. This is very different from the typical use, where a sum of money is transferred from an individual to an organization without personal contact. If this were an intuited example, I might judge it grammatical (at least marginally so) for similar reasons, while another researcher, unaware of my subtle reconceptualization, would judge it ungrammatical, leading to no insights whatsoever into the semantics of the verb *donate* or the valency patterns it occurs in.

1.3 Intuition data vs. corpus data

As the preceding section has shown, intuited judgments are just as vulnerable as corpus data as far as the major points of criticism leveled at the latter are concerned. In fact, I have tried to argue that they are, in some respects, more vulnerable to these criticisms. For those readers who are not yet convinced of the need for corpus data, let me compare the quality of intuited "data" and corpus data in terms of two aspects that are considered much more crucial in methodological discussions outside of linguistics than those discussed above:

1. data reliability (roughly, how sure can we be that other people will arrive at the same set of data using the same procedures);
2. data validity or epistemological status of the data (roughly, how well do we understand what real world phenomenon the data correspond to);[5]

As to the first criterion, note that the problem is not that intuition "data" are necessarily wrong. Very often, intuitive judgments turn out to agree very well with more objective kinds of evidence, and this should not come as a surprise. After all, as native speakers of a language, or even as advanced foreign-language speakers, we have considerable experience with using that language actively (speaking and writing) and passively (listening and reading). It would thus be surprising if we were categorically unable to make statements about the probability of occurrence of a particular expression.

Instead, the problem is that we have no way of determining introspectively whether a particular piece of intuited "data" is correct or not. To decide this, we need objective evidence, obtained either by serious experiments (including elicitation experiments) or by corpus-linguistic methods. But if that is the case, the question is why we need intuition "data" in the first place. In other words, intuition "data" are simply not reliable.

The second criterion provides an even more important argument, perhaps *the* most important argument, against the practice of intuiting. Note that even if we manage to solve the problem of reliability (as systematic elicitation from a representative sample of speakers does to some extent), the epistemological status of intuitive data remains completely unclear. This is particularly evident in the

[5] Readers who are well-versed in methodological issues are asked to excuse this somewhat abbreviated use of the term *validity*; there are, of course, a range of uses in the philosophy of science and methodological theory for the term validity (we will encounter a different use from the one here in Chapters 2.3 and 4).

case of grammaticality judgments: we simply do not know what it means to say that a sentence is "grammatical" or "ungrammatical", i.e., whether grammaticality is a property of natural languages or their mental representations in the first place. It is not entirely implausible to doubt this (cf. Sampson 1987), and even if one does not, one would have to offer a theoretically well-founded definition of what grammaticality is and one would have to show how it is measured by grammaticality judgments. Neither task has been satisfactorily undertaken.

In contrast, the epistemological status of a corpus datum is crystal clear: it is (a graphemic representation of) something that a specific speaker has said or written on a specific occasion in a specific situation. Statements that go beyond a specific speaker, a specific occasion or a specific situation must, of course, be inferred from these data; this is difficult and there is a constant risk that we get it wrong. However, inferring general principles from specific cases is one of the central tasks of all scientific research and the history of any discipline is full of inferences that turned out to be wrong. Intuited data may create the illusion that we can jump to generalizations directly and without the risk of errors. The fact that corpus data do not allow us to maintain this illusion does not make them inferior to intuition, it makes them superior. More importantly, it makes them normal observational data, no different from observational data in any other discipline.

To put it bluntly, then, intuition "data" are less reliable and less valid than corpus data, and they are just as incomplete and in need of interpretation. Does this mean that intuition "data" should be banned completely from linguistics? The answer is no, but not straightforwardly.

On the one hand, we would deprive ourselves of a potentially very rich source of information by dogmatically abandoning the use of our linguistic intuition (native-speaker or not). On the other hand, given the unreliability and questionable epistemological status of intuition data, we cannot simply use them, as some corpus linguists suggest (e.g. McEnery & Wilson 2001: 19), to augment our corpus data. The problem is that any mixed data set (i.e. any set containing both corpus data and intuition "data") will only be as valid, reliable, and complete as the weakest subset of data it contains. We have already established that intuition "data" and corpus data are both incomplete, thus a mixed set will still be incomplete (albeit perhaps less incomplete than a pure set), so nothing much is gained. Instead, the mixed set will simply inherit the lack of validity and reliability from the intuition "data", and thus its quality will actually be *lowered* by the inclusion of these.

The solution to this problem is quite simple. While intuited information about linguistic patterns fails to meet even the most basic requirements for scientific

data, it meets every requirement for scientific *hypotheses*. A hypothesis has to be neither reliable, nor valid (in the sense of the term used here), nor complete. In fact, these words do not have any meaning if we apply them to hypotheses – the only requirement a hypothesis must meet is that of *testability* (as discussed further Chapter 3). There is nothing wrong with introspectively accessing our experience as a native speaker of a language (or a non-native one at that), provided we treat the results of our introspection as hypotheses about the meaning or probability of occurrence rather than as facts.

Since there are no standards of purity for hypotheses, it is also unproblematic to mix intuition and corpus data in order to come up with more fine-grained hypotheses (cf. in this context Aston & Burnard 1998: 143), as long as we then *test* our hypothesis on a pure data set that does not include the corpus-data used in generating it in the first place.

1.4 Corpus data in other sub-disciplines of linguistics

Before we conclude our discussion of the supposed weaknesses of corpus data and the supposed strengths of intuited judgments, it should be pointed out that this discussion is limited largely to the field of grammatical theory. This in itself would be surprising if intuited judgments were indeed superior to corpus evidence: after all, the distinction between linguistic behavior and linguistic knowledge is potentially relevant in other areas of linguistic inquiry, too. Yet, no other sub-discipline of linguistics has attempted to make a strong case against observation and for intuited "data".

In some cases, we could argue that this is due to the fact that intuited judgments are simply not available. In language acquisition or in historical linguistics, for example, researchers could not use their intuition even if they wanted to, since not even the most fervent defendants of intuited judgments would want to argue that speakers have meaningful intuitions about earlier stages of their own linguistic competence or their native language as a whole. For language acquisition research, corpus data and, to a certain extent, psycholinguistic experiments are the only sources of data available, and historical linguists must rely completely on textual evidence.

In dialectology and sociolinguistics, however, the situation is slightly different: at least those researchers whose linguistic repertoire encompasses more than one dialect or sociolect (which is not at all unusual), could, in principle, attempt to use intuition data to investigate regional or social variation. To my knowledge, however, nobody has attempted to do this. There are, of course, descriptions of

individual dialects that are based on introspective data – the description of the grammar of African-American English in Green (2002) is an impressive example. But in the study of actual *variation*, systematically collected survey data (e.g. Labov et al. 2006) and corpus data in conjunction with multivariate statistics (e.g. Tagliamonte 2006) were considered the natural choice of data long before their potential was recognized in other areas of linguistics.

The same is true of conversation and discourse analysis. One could theoretically argue that our knowledge of our native language encompasses knowledge about the structure of discourse and that this knowledge should be accessible to introspection in the same way as our knowledge of grammar. However, again, no conversation or discourse analyst has ever actually taken this line of argumentation, relying instead on authentic usage data.[6]

Even lexicographers, who could theoretically base their descriptions of the meaning and grammatical behavior of words entirely on the introspectively accessed knowledge of their native language have not generally done so. Beginning with the Oxford English Dictionary (OED), dictionary entries have been based at least in part on *citations* – authentic usage examples of the word in question (see Chapter 2).

If the incompleteness of linguistic corpora or the fact that corpus data have to be interpreted were serious arguments against their use, these sub-disciplines of linguistics should not exist, or at least, they should not have yielded any useful insights into the nature of language change, language acquisition, language variation, the structure of linguistic interactions or the lexicon. Yet all of these disciplines have, in fact, yielded insightful descriptive and explanatory models of their respective research objects.

The question remains, then, why grammatical theory is the only sub-discipline of linguistics whose practitioners have rejected the common practice of building models of underlying principles on careful analyses of observable phenomena. If I were willing to speculate, I would consider the possibility that the rejection of corpora and corpus-linguistic methods in (some schools of) grammatical theorizing are based mostly on a desire to avoid having to deal with actual data, which *are* messy, incomplete and often frustrating, and that the arguments against the use of such data are, essentially, post-hoc rationalizations. But whatever the case

[6]Perhaps Speech Act Theory could be seen as an attempt at discourse analysis on the basis of intuition data: its claims are often based on short snippets of invented conversations. The difference between intuition data and authentic usage data is nicely demonstrated by the contrast between the relatively broad but superficial view of linguistic interaction found in philosophical pragmatics and the rich and detailed view of linguistic interaction found in Conversation Analysis (e.g. Sacks et al. 1974, Sacks 1992) and other discourse-analytic traditions.

may be, we will, at this point, simply stop worrying about the wholesale rejection of corpus linguistics by some researchers until the time that they come up with a convincing argument for this rejection, and turn to a question more pertinent to this book: what exactly constitutes corpus linguistics?