INTERNATIONAL MULTIDISCIPLINARY JOURNAL FOR RESEARCH & DEVELOPMENT
SJIF 2019: 5.222 2020: 5.552 2021: 5.637 2022:5.479 2023:6.563 2024: 7,805
eISSN :2394-6334 https://www.ijmrd.in/index.php/imjrd Volume 11, issue 05 (2024)

UDK:81-133

# THE NEED FOR CORPUS DATA

**Kaxorova Nargiza Nusratovna**

The teacher of Bukhara State university

Kaxorovanargiza5@gmail.com

**Abstract:** 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.

**Key words:** Corpus, crudely, a large collection of authentic text, genuine communicative situations, corpus linguistics.

## Introduction

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 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".

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;

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.

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 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.

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. (avonwildlife).

## References

1. Altenberg, Bengt. 1980. Binominal NP’s in a thematic perspective: Genitive vs. of–constructions in 17th century English. In Sven Jacobsen (ed.), Papers from the Scandinavian Symposium on Syntactic Variation (Stockholm Studies in English 52), 149–172.
2. Stockholm: Almqvist & Wiksell. APA, American Psychiatric Association. 2000. Diagnostic and statistical manual of mental disorders: DSM-IV-TR. 4th ed., text revision. Washington, DC: American Psychiatric Association. Aston, Guy & Lou Burnard. 1998.
3. The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press. Baayen, Harald. 2008.
4. Analyzing linguistic data: A practical introduction. Cambridge & New York: Cambridge University Press. Baayen, Harald. 2009.
5. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), Corpus linguistics: An international handbook, vol. 2 (Handbooks of Linguistics and Communication Science 29), 899–919.
6. Berlin & New York: De Gruyter Mouton. Baker, Carolyn D. & Peter Freebody. 1989.

7. Children's first school books: Introductions to the culture of literacy. Oxford & Cambridge, MA: Blackwell. Baker, Paul. 2010a.

8. Sociolinguistics and corpus linguistics (Edinburgh sociolinguistics). Edinburgh: Edinburgh University Press.