The Essential Contributions of Corpora in Language Research

Hamza Cherifi
Khemis Miliana University, Algeria

Abstract: A major, overriding outcome of the organic complexity of language, which legitimizes the rival claims inherent in its description, is the empiricist/rationalist “paradigm war” over sources of language data. While rationalists advance the pervasiveness of competence and, by extension, the reliability of introspective methodology as compared to the “skewed” performance data, empiricist accounts treat naturally occurring language not only as dependable but amenable to a wider scope of investigation as well. However, the inception of Corpus Linguistics as a definable approach in the 1980s marked the start of a new exploratory potential that surpassed and identified drawbacks in the then-popular Chomsky’s linguistics. A corpus-based approach to language study markedly differs from their ‘rivals’, not only as the former employs a set of identifiable research methods but, most importantly, as using a corpus allows for a breadth of coverage that renders possible addressing frequency–based questions. This property enables both the objective verification of introspection-based assumptions and an ongoing reappraisal of existing descriptions. This paper probes the nature of corpus linguistics as a methodology to language study by elaborating on core tenets of corpus-based approach, and by assessing repercussions of the breadth-of-coverage property on a variety of language-related areas to examine what insights and contributions corpora may bring in each.

Key Words: Corpus Linguistics, Language Study, Frequency Analysis, Variety, Lexicography, Discourse Analysis
1. Background to Corpus Linguistics

A linguistic investigation—whether scientific or otherwise—demands a relative specification of the approaches and methods underpinning the study. Corpus Linguistics represents a principled approach to the investigation of linguistic phenomena. The domain marks a turning point in language research as it opens up the scale of researchability, allowing for grounded exploration of uncharted areas. It does so through specific methods of data collection, analysis and interpretation said to be ostensive markers of a corpus-based language enquiry. Yet, before we venture on a discussion of these principles, we need to address the much more basic question: What is Corpus Linguistics?

1.1 Early Corpus Linguistics

In the most profound sense, corpus linguistics is the area in linguistic research concerned with the study of language in use, with the other side of the continuum being intuition, or the introspection of how language is used. The method utilizes analysis of large collections of language data. Prior to the paradigm shift Noam Chomsky has sparked in the perception of language data, and hence in the orientations of linguistic enquiry, research into language can be perfectly described as corpus-based, for the essence of a corpus-based approach—putting collections of large language data under study—did prevail with the huge undertake of manual assemblage and analysis. The approach, however, was underlined by a serious fallacy: That language is finite and, thus, amenable to collection and enumeration. This drawback lined up with Chomsky’s case against empiricist accounts to discredit corpus-based methodology, which has undergone a period of unpopularity.

1.2 What Chomsky Said

Central to Chomsky’s linguistics is the view of language being an inborn (endowed), mentally modular entity independent of social and environmental variables. This cognitive explicandum led Chomsky to perceive performance, the actual use of language, as underlined by competence, the abstract, pervasive set of internalized rules which, as such, better explains the speaker’s knowledge of language. A corpus, by its very nature, marks performance data, which do not reliably tell of the speaker’s knowledge of language, especially that factors like anxiety and memory limitation exert a great deal of influence on its state.

For Chomsky and his proponents, the finite set of rules composing syntactic competence allows the generation of non-enumerable sets of phrases in performance. Thus, corpus linguists bother searching in an immense number of words for facts that may be available just by “sitting and thinking for few minutes”—introspection. Consider the following reported dialogue between Chomsky and Hatcher:

Chomsky: The verb perform cannot be used with mass word objects: One can perform a task but one cannot perform labour.

Hatcher: How do you know if you do not use the corpus and have not studied the verb perform?

Chomsky: How do I know? Because I am a native speaker of the English language(emphasis added).

(Hill, 1962, p. 29, cited in McEnery and Wilson, 2001, p. 11)

Chomsky’s manifesto sees no other than the linguist (the analyst) or the native speaker as the sole explicandum of linguistics. In a way, Chomsky’s pronouncement does justice to the utility of introspective data which, in language study, can satisfy questions corpus data fail
to resolve. It is only by eliciting the opinion of a native speaker or an expert user of the language that we can differentiate ungrammatical utterances from those which are grammatical but have not occurred (McEnery and Wilson, 2001, p. 14).

1.3 Why Corpus Linguistics Persisted

Introspective data cannot compare to the actual usage of language. This follows that, unlike unconscious language use, introspection requires a process of consciousness which comes with specific functioning:

One of the most peculiar features of consciousness is its narrowness compared to the vast and almost limitless unconscious processing capacity at our disposal. Consciousness is “often seen to … focus the brain’s vast unconscious resources and make them available for specific cognitive use (Dörnyei, 2009, p. 133).

Similarly, Schmidt (1990, p. 30) comments that “people learn much about the things they attend to and do not learn much about the things they do not attend to”. By extension, introspection is limited in its exploratory potential as areas grouped under the heading of linguistics “could not effectively be studied in the artificial world of well-formedness and idealized speaker-hearer competence” (McEnery & Wilson, 2001, p. 14). For their availability, introspective data condition a metalinguistic awareness needed to report thought processing. Observing language data remains the dominant method for studying language variation and second language acquisition. And “if you doubt so, try asking an eighteenth-month child in a one-word stage whether the word *mow-cow* is a verb or a noun” (McEnery & Wilson, 2001, p. 14).

Introspective data can be highly biased and skewed as the analyst tends to provide one-sided views likely to be shaped by one’s dialect. It is impossible for the non-corpus linguist (the armchair linguist) to access and comment on other uses of language in various contexts. A corpus can encompass the actual language of many users in different modes and within many registers. Also, while introspective data cannot be objectively verified, the corpus provides a means to check the interpretive validity of the statements by rendering the analysis visible through annotation and concordance. Figure 1 below shows a sample of concordance for the uses of *and* by learners with relative proficiency as transcribed in the Michigan Corpus of Academic Spoken English.

Figure 1: Concordance in the Michigan Corpus of Academic Spoken English.
Non-corpus linguists are constrained in terms of frequency-based information. As McEnery and Wilson (2001) describes, “human beings have the vaguest notion of the frequency of a construction or word” (p. 15). Corpora are the only source of quantitative information.

1.4 Properties of a Corpus

Etymologically, the word corpus is the Latin for body, which implies that it is the large kind of text that qualifies as corpus. However, as a unit of linguistic analysis, a corpus differs in fundamentally several accounts from the examination of texts. McEnery and Wilson characterize a corpus as having the following properties:

- **Representativeness:** In linguistics, we concern ourselves with whole varieties of language than with texts of individual authors. Since it is impossible to analyze the whole utterances of a given variety, we resort to a sample of language that is more or less representative of that variety. This representativeness is one of the essential properties of the corpus as a text. In building a corpus of a particular language variety, we need to make it as representative as possible so that we are with a maximally reliable picture of the tendencies and frequencies of the data under examination. For the corpus of mid-nineteenth century English to be representative, it would not suffice to reduce our investigation to Charlotte Bronte’s writings, nor would it be representative to confine our analysis in the genre of literature.

- **Finite size:** A corpus is not a random collection of an open-ended set of words. It has a finite set of words (the grand number) at which the collection stops. The British National Corpus, for instance, has the grand number of 100,000,000 words. An exception is the monitor corpus, which is an open-ended entity to which texts are constantly added. Such corpora are used for lexicographic works looking at new occurrences or at the changing meaning of words.

- **Machine-readable:** For many years the term ‘corpus’ has been used to merely mean printed texts. Nowadays, the term implies that the body of text is machine-readable. The advantage of computerized corpora is that they can be easily manipulated for diverse research objectives. In printed corpora, it would require a cover-to-cover search to study the frequency of, for instance, the word ‘book’. Thanks to concordance software, such tasks may take few seconds.

- **Standard reference:** Generally, corpora come with a standard reference that is indicative of the variety it represents. To illustrate, there exist the Brow Corpus of Written American English, the LOB Corpus of Written British English and the London-Lund Corpus of Spoken British English.

2. Corpora in Language Study

As a corpus makes available a huge number of language data for analysis, we can afford to make reliable claims about preferred patterns and choices, and to characterize a given language variety, text type or genre as having a unique set of choices as part of a distinctive typology. These judgments are based on frequency evidence: A lexical or structural choice is unique to a genre or variety because it occurs more frequently in that variety and not in another. The frequent-infrequent distinction is used to make other distinctions relevant to the study of a variety of language areas. In what follows, we examine how the frequent-infrequent affordance gets deployed to address language-related questions. We shall also discuss the potential of concordance, namely in discourse analysis and lexicography.
2.1 Corpora in Language Teaching

Although the division between empiricists and rationalists features in the examples used in textbooks, the reason why invented examples are still used may not be due to allegiance to intuition-based incentive. Instead, the preference for invented samples of language may follow from perceiving naturally occurring language as improper for the context of the classroom, which requires a type of language less bound up with culture-specific and immersion situations. It could also be that the language meant to be taught and learnt is other than communicative and genuine. Where this is the aim, corpora are valuable resources for familiarizing students with reading tangible texts and for modeling the language of real communicative encounters.

McEnery and Wilson reviewed several studies comparing constructions in textbooks with the constructions in standard corpora to examine some differences and to gauge the authenticity of examples. Also, frequency analysis verifies claims about common usage as presented in textbooks. One further possibility could be that the very language teaching textbooks get used as corpora meant for analyzing and assessing ESL material.

Learner corpora are another way in which corpus linguistics helps research into language teaching. This type of corpora is made up of the language produced by its learners, offering an incentive for studying and characterizing interlanguage, the learner’s typical language. Potentially, learner corpora are suitable most for error analysis studies seeking to construe a profile of learners’ errors. This does tell of difficulties encountered by learners, and hence findings of error analysis can be used for remedial teaching or in an in-advance focus on difficult language areas.

2.2 Corpora in Semantics

Semanticists take as their primary objective the disentangling of the meaning of individual words. In so doing, a word gets defined by relating its meaning to the meaning of other words, and sense is the result of recognizing the relationship between words. For instance, the *Oxford Advanced Learner’s Dictionary* (2010, 8th ed.) defines the word *boar* as “a male pig which has not been castrated”. Clearly, the definition locates the term within a general concept of *pig* followed by the rest of specifications. *Boar and hog* are subdivisions, or hyponyms, of the general superordinate *pig*. By the same token, word meaning can be explained measuring it against the other set of sense relations, including synonyms, near-synonyms, antonyms, etc. (see Palmer, 1981). Likelihood of overlap and fluctuation between the entry defined and the items used to define it, presupposes precision. Hence, it is fundamental to properly identify the conceptual saliencies of words. Here, corpora assist in offering an objective, principled measure for ascribing denotations to linguistic items, instead of relying on a linguist’s introspection or intuition. A word may have several meanings or may be used to denote an associative or emotive signification since no two synonyms are exactly interchangeable. Concordance offers a valuable incentive for these dynamics. Also, frequency analysis helps to establish a hierarchy, which allows a word to be defined by others that are more familiar to readers.

Another main contribution of corpora in semantics is that it makes the notion of *fuzzy categories* more principled. In defining the meaning of words, semanticists tend to first locate an item within a general category. The items *boar* and *hog* fall within the semantic type of *pig*. While the spreading out of items make categories appear hard and distinctive, McEnery and Wilson remark that “psychological works on categorizations have suggested that typically
cognitive categories are not hard and fast so that it is not so much a question of whether or not a given item belongs in a given category as of how often it falls within that category as opposed to another” (McEnery and Wilson, 2001, p. 113). This has important applications in the link between semantic boundaries and corpus linguistics. Analyzing naturally occurring language, we are not concerned with whether or not an item occurs within a given category; we are rather preoccupied with comparing the occurrence of this item along several categories containing the item in relative scales. Hence it is more a matter of frequency of inclusion instead of mere inclusion: Whether an item is frequent or not marks the criterion gauging its belonging to a category. To demonstrate the importance of corpora in establishing this gradation, McEnery and Wilson reviewed Mindt’s (1991) study of the subjects of verbs with future time indicator. While previous theoretical frameworks would have shoehorned one construction to one category, Mindt’s study revealed that personal subjects occur most with the present progressive whereas the simple future has the lowest association with human agency. We can by the same token establish gradation in ascribing items to macro-categories as texts and genre. To illustrate, let us verify whether the verb believe is discarded in academic discourse as it implicates value judgment not supported by evidence. Figure 2 shows results for the frequency of the verb believe across different genres:

![Figure 2: Distribution of the verb believe in the Corpus of Contemporary American English.](image)

Results seem to validate the oft-repeated conviction that the phrase believe is discarded in academic writing as the verb comes with the lowest frequency in the academic category. The categories ‘Fiction’ and ‘Spoken’ contains the highest frequencies of the verb. This follows that the discourse community in both genres do not pose the restrictions known to academic writing. This example shows the benefits of corpora when relating items to categories.

At another level, collocational behavior implicates the notion of choice, where a given combination of words may be either familiar or unfamiliar to recipients, who generally build
expectation of usage on a presumption established due to recurrent, consensually exchanged constructions. Where the combinations are not very much anticipated, they still function as prototypes for perceiving the picture meant by violating the norms. In this sense a corpus helps define what combinations are normative or conventional. Again, the notion of frequency solves the disputes over the appropriateness of language, making conventionality and acceptability bound up with the neutral criteria of recurrence rather than suspended or related to individuals with subjective claims about conventions. Figure 3 displays the subject collocates of the verb *abate*.

![Figure 3: Common subjects of the verb abate in the Corpus of Contemporary American English.](image)

At the same time, the very scale of infrequent collocation enjoys ‘communicative’ relevance in other genres. What is held as idiosyncratic collocation in one genre may be regarded as creative metaphor meant to spark a specific effect. For instance, the phrase *my father abated after screaming a lot at me* makes up a metaphor used to show how angry a person is: Using *abate* with a person subject serves allocating this person the properties of inanimate subject. *Abate* usually occurs with the likes of *results, energy, and electricity*. Hence *my father abated* instead of *my father calmed* is intended to portray a person as mindless, violent or as having no self-control.

### 2.3 Corpora in Lexicography

Lexicographers and dictionary makers set out to account for the usage of terms in different contexts and at different times, and hence the practice of citing examples from naturally occurring language preceded the inception of corpus linguistics. Yet, corpora have been changing the manner in which lexicographers treat and make use of language data: “Corpora…now mean that a lexicographer can sit at a computer terminal and call up the examples of the usage of a word or phrase from many millions of words of texts in few seconds” (McEnery and Wilson, 2001, p. 106). This implies that information about usage can
be revised and made more comprehensive, making definitions more precise. Adding to this, the large and constantly collected sample of data allows capturing up-to-date usage of terms, thereby accounting for the dynamics and change of word meaning. McEnery and Wilson went on to show how corpora give insights for morphology studies. Here, corpora assist in linking a given manner of structuring words to an inclusive category. It is the frequency analysis that helps establish, for instance, how different varieties of English—not only British or American—formulate the past tense (e.g. learn to learned), or whether contractions are really less recurrent in academic genre.

2.4 Corpora in Speech

A spoken text has its own phonological and prosodic features as part of its text typology. An academic talk arguably shows prosodic features different from those found in a political speech or in a liturgy. Taking a large sample of data means that generalizations about the features of spoken texts can be made, and the sub-corpora representing variant genres or any other speech variables offer a platform for within-corpus comparisons. Spoken corpora normally operate by prosodic annotations used to tag the transcribed text. McEnery and Wilson conclude that research using corpora tends to investigate the relatedness between the prosodic features of a text and its linguistic properties, for instance whether the intonation group boundary associates with a given syntactic marker. A serious problem with speech corpora lies in the very transcription of speech. Looking at prosodic features making a corpus of sounds, one might raise the question of how far the transcription reflects the actual speech. Transcribers may have different perception of, for instance, pitch height, resulting in discrepant identification of fall and raise tones.

2.5 Corpora in Text Linguistics

One major area in which textual properties are distinguished appears the distinction between spoken and written forms of communication in order to attain the properties salient of different communicative channels. However, as corpora provides a set of text belonging to several types and genres, researchers have used corpora to identify the textual similarities and differences between these larger linguistic units. In so doing, a considerable set of text linguistics studies identified textual variation using Halliday and Hasan’s (1976) model of cohesion. Hence, a text or genre is characterized with reference to the nature of reference and dependency relations holding between the lexico-grammatical elements.

Let us consider Kaltenböck’s (2004) examination of non-extraposition clause to illustrate the investigatory potential of corpora and how information is structured and ordered across texts. Kaltenböck purported to investigate the status of non-extraposition clause in written and spoken text as compiled in the International Corpus of English. Non-extraposition clause denotes a case in which an entire clause takes subject position. This template runs counter to conventional terms: We hardly think of a subject as something other than a noun phrase. Kaltenböck explains the non-extraposition clause in relation to its it-extraposition counterpart:

<table>
<thead>
<tr>
<th>Non extraposition clause</th>
<th>It-extraposition clause</th>
</tr>
</thead>
<tbody>
<tr>
<td>That John will go to Paris is surprising</td>
<td>It is surprising that John will go to Paris</td>
</tr>
<tr>
<td>What John will do in Paris is obvious</td>
<td>It is obvious what John will do in Paris</td>
</tr>
<tr>
<td>Living in Paris is a fun</td>
<td>It is a fun living in Paris</td>
</tr>
<tr>
<td>To find a job is imperative</td>
<td>It is imperative to find a job</td>
</tr>
</tbody>
</table>
Although the two clauses share the same lexico-syntactic properties, the fact that they are not exchangeable in texts shows that they serve different functional purposes. Kaltenböck reviews that “this functional difference is already indicated by frequency evidence, which shows that non-extraposition occurs less frequently than its extraposited counterpart and therefore has to be taken as the satisfactorily marked construction of the two” (p. 220). The following uses of and distributions of non-extraposition clause were identified:

- Non-extraposition clause is outnumbered by its fit-extraposition counterpart.
- The ing- and wh-clauses are more frequent than their extraposited counterparts.
- The non-extraposition clause occurs in the written category as twice as in the spoken one.
- Non-extraposition clause has a cohesive function as it carries old information.
- Non-extraposition clause carries new information in situations where a new topic is introduced to the discourse as if it were generally known.
- Non-extraposition clause characterizes persuasive writing than other text types.
- Frequency of non-extraposition clause correlates with formality.
- Frequency of non-extraposition clause matches the style in newspaper articles and in press editorials.

Operating within the same track of using corpora in characterizing information distribution in texts and genres, Bruti (2004) examines the uses and functions of cataphoric reference. Unlike anaphoric reference, where an item stands for an identification situated before, cataphoric relations involve situating the referring tools before their identification, which gets resolved by reading forward. Hence, this strategy of delaying referential resolution instead of making the identification already known characterizes texts and genres in relative scales.

2.6 Corpora in Pragmatics and Discourse Analysis

Before we establish the relevance of corpus linguistics in research within these two interrelated tracks, we first expose their scope and principles to examine how their concerns line up with the affordance of concordance. In so doing, we take from Coulthard’s (1985) introduction. Both pragmatics and discourse analysis embody specific approaches to describe language. While syntax looks at how words are arranged to form sentences and clauses by identifying the functions of grammatical categories, and while semantics targets the common identification of isolated words and sentences, pragmatics and discourse are disciplines concerned with the context of the utterance and how meaning can be explained by knowledge of the physical world or shared knowledge. Both disciplines consider interaction as the manifestation of meaning, by looking at ways in which interlocutors can convey information more than the words they deploy. Both areas seek to explicate how a given stretch of words is meaningful in a given setting (felicity conditions).

As pragmatics and discourse analysis deal with context, the latter is represented in three aspects of the physical and psychological parameters underlying the relations that hold the speaker and the hearer. Situational context, or the physical world holding the interaction, stands for the meaning that the surrounding object and gestures can contribute to the overall interpretation of the speaker’s utterance. The second type of context is that of background knowledge, which determines the nature of meaning negotiated and the understandings interactants share and take for granted. This knowledge could be categorized as cultural, where knowledge about particular realities identify interlocutors as belonging to the same
speech community (*shared schemata*), or interpersonal, referring to the fact interlocutors share a very specific schematic reference that they draw on. Generally, the type of talk that involves shared cultural knowledge displays shared assumption about certain realities. It is the type of talk in which speakers can be seen as accommodating their attitudes to be acceptable and to be seen as belonging to the same group. The third type of context is the co-text, which embodies what speakers know about what they have been saying. The identification of what has been said is done through perceiving the nature of reference, which is either *exophoric*, where the referent is outside the actual text, or *endophoric*, with the referent either preceding or proceeding its identification.

This interest in the function of utterance in context escalated into *The Speech Act Theory*, which suggests that utterances perform acts analyzable in terms of three dimensions. The first level denotes what the utterance is about in terms of syntax and semantics, making the *locutionary act*. The second level of analysis concerns the act that the utterance performs in real world—*illocutionary act*. The third level looks at the impact of the act on the recipient—*perlocutionary act*. Speech acts might be embedded in utterances containing *performatives*: verbs that perform the speech act.

In terms of frequency analysis, this very practice of ascribing given manners of speech act performance to levels of formality or to specific communicative encounters needs to be subsided with frequency argument: A given manner of performing a speech act is said to be typical to a genre because it is more frequent in that genre compared to the rest. At another scale, because degrees of directness are connected, whether the semantics of the utterance is present, frequency analysis can tell the level of indirectness across several genres and along several communicative situations. In the measure where directness and bold strategies are held to correlate with how much threat is caused to the face of interactants, corpus linguistics could also refine claims about politeness by making it a matter of frequency.

To illustrate the benefits of corpora in accounting for the functions of utterances, consider Diani’s (2004) study of the functions of *I don’t know* in English searching the Birmingham Spoken Corpus. The utterance has as its semantic function “a declaration of insufficient knowledge”. When used in context, the utterance is found to serve the following functions illustrated by Diani from the examined corpus:

**To show disagreement when prefaced by well:**
A: They’re let out again and they are doing it again.
B: Well I don’t know. Again we are guessing.

**To mitigate talk in order to preserve the interlocutor’s face through displaying reluctance:**
A: They don’t cost anymore than any other car parks.
B: Oh I don’t know. I mean where I park in my little very cheap ten pence an hour.

**To ask for clarification through displaying uncertainty:**
A: Like what?
B: Oh I don’t know somebody who is more needy than people spending money.
A: Yeah.

**To display search for better ways of expressions:**
I mean I don’t know. I just feel happy wearing what I do.

**To support one’s argument:**
I do not know. I mean I think that we have there a lot of groups.
Working within the same tradition, Bamford (2004) studies the gestural and symbolic uses of the deictic *here* in a corpus of academic lectures. The large sample of naturally occurring language unfolds the uses of words and phrases, making it possible to account for the functions of specific stretch of language. This identification gets assimilated to the entire description of words and phrases.

**Conclusion**

We have seen how addressing a wide range of language related questions can be achieved by means of corpora. As demonstrated, frequency analysis informs of what is salient and conventional, and thus assists in the characterization of typologies (semantic, phonological, textual, etc.). The immense set of language data and concordance helps catering for the usage and functions of words and phrases. It should be noted that resorts to corpora might be motivated by verification purposes: Whether and what we know about how language is prototypically used—either by intuition or by inherited claims—can be supported by an empirical indicator. It is well to remember that even though the collection and design of corpora is hardly random, the process needs to be guided by a well-defined scope of researchability. This follows that language researchers may not encompass the entire benefits of corpus tools. It is likely that a clear demonstration of the uses of a given corpus would stimulate further inquisitiveness. Corpus designers need to demonstrate how a query serves specific research questions so that users understand what could possibly be done with all the collected data and with all the research options available in corpus software. Within the realm of the language teaching enterprise, for instance, corpus designers would have to illustrate the relevance of available data to language pedagogy. This can be done by indicating that the corpus can answer questions like what features are frequent most in the variety being taught in class; what differences are there between written and spoken language; whether the learners’ output is conventional or idiosyncratic; what characterizes learners’ language most; what common errors are made by learners; what features of negative language transfer are present most; what cultural elements are present in learners’ language, etc. It should be reminded that no matter how wide the scope of investigation might be, corpus linguistics remains a methodology for handling the performance type of language data, which only escalates the empiricist epistemology.
References


Acknowledgement

I owe the motivation for this work to my teacher, Bel Abbes Neddar.