BEST JOURNAL OF INNOVATION IN SCIENCE, RESEARCH AND DEVELOPMENT
ISSN: 2835-3579
Volume:3 Issue:6 | 2024
www.bjisrd.com

Corpus Linguistics as a Scientific Method

Kaxorova Nargiza Nusratovna,
The teacher of Bukhara State University

***Abstract*:** *The first way entails a relatively open-minded approach to our data. We might have some general expectation of what we will find, but we would put them aside and simply start collecting observations and look for patterns. If we find such patterns, we might use them to propose a provisional generalization, which we successively confirm, modify or replace on the basis of additional observations until we are satisfied that we have found the broadest generalization that our data will allow – this will then be the answer to our research question.*

Key words: *disrepute, comeback, reliable, generalizations, formulate, correlations, supplementary, lexicology, diachronic perspectives.*

Introduction

This so-called inductive approach was famously rejected by the Austrian-British philosopher Karl Popper for reasons that will become clear below, but after a period of disrepute it has been making a strong comeback in many disciplines in recent years due to the increasing availability of massive amounts of data and of tools that can search for correlations in these data within a reasonable time frame (think of the current buzz word “big data”).

Such massive amounts of data allow us to take an extremely inductive approach – essentially just asking “What relationships exist in my data?” – and still arrive at reliable generalizations. Of course, matters are somewhat more complex, since, as discussed at the end of the previous chapter, theoretical constructs cannot directly be read off our data. But the fact remains that, used in the right way, inductive research designs have their applications. In corpus linguistics, large amounts of data have been available for some time (as mentioned in the previous chapter, the size even of corpora striving for some kind of balance is approaching half-a-billion words), and inductive approaches are used routinely and with insightful consequences (Sinclair 1991 is an excellent example). The second way of stating research questions entails a more focused way of approaching our data. We state our hypothesis before looking at any data, and then limit our observations just to those that will help us determine the

truth of this hypothesis (which is far from trivial, as we will see presently). This so-called deductive approach is generally seen as the standard way of conducting research (at least ideally – actual research by actual people tends to be a bit messier even conceptually). We will generally take a deductive approach in this book, but it will frequently include inductive (exploratory) excursions, as induction is a often useful in itself (for example, in situations where we do not know enough to state a useful working hypothesis or where our aim is mainly descriptive) or in the context of deductive research (where a first exploratory phase might involve inductive research as a way of generating hypotheses).

There is only one way: we have to find the word in question. We could, for example, describe the concept Forward-Facing Window of Car to a native speaker or show them a picture of one, and ask them what it is called (a method used in traditional dialectology and field linguistics). Or we could search a corpus for all passages mentioning cars and hope that one of them mentions the forward-facing window; alternatively, we could search for grammatical contexts in which we might expect the word to be used, such as ⟨ through the NOUN of POSS.PRON car ⟩ (see Section 4.1 in Chapter 4 on how such a query would have to be constructed). Or we could check whether other people have already found the word, for example by searching the definitions of an electronic dictionary. If we find a word referring to the forward-facing window of a car, we have thereby proven its existence – we have verified the statement in (1). But how could we falsify the statement, i.e., how could we prove that English does not have a word for the forward-facing window of a car? The answer is simple: we can't. As discussed extensively in Chapter 1, both native-speaker knowledge and corpora are necessarily finite. Thus, if we ask a speaker to tell us what the forward-facing window of car is called and they don't know, this may be because there is no such word, or because they do not know this word (for example, because they are deeply uninterested in cars). If we do not find a word in our corpus, this may be because there is no such word in English, or because the word just happens to be absent from our corpus, or because it does occur in the corpus but we missed it. If we do not find a word in our dictionary, this may be because there is no such word, or because the dictionary-makers failed to include it, or because we missed it (for example, because the definition is phrased so oddly that we did not think to look for it – as in the Oxford English Dictionary, which defines windscreen somewhat quaintly as "a screen for protection from the wind, now esp. in front of the driver's seat on a motor-car" (OED, sv. windscreen)). No matter how extensively we have searched for something (e.g. a word for a particular concept), the fact that we have not found it does not mean that it does not exist.

The statement in (1) is a so-called "existential statement" (it could be rephrased as "There exists at least one x such that x is a word of English and x refers to the forward-facing window of a car"). Existential statements can (potentially) be verified, but they can never be falsified. Their verifiability depends on a crucial condition hinted at above: that all words used in the statement refer to entities that actually exist and that we agree on what these entities are. Put simply, the statement in (1) rests on a number of additional existential statements, such as "Languages exist", "Words exist", "At least one language has words", "Words refer to things", "English is a language", etc. There are research questions that take the form of existential statements. For example, in 2016 the astronomers Konstantin Batygin and Michael E. Brown proposed the existence of a ninth planet (tenth, if you cannot let go of Pluto) in our solar system (Batygin & Brown 2016). The existence of such a planet would explain certain apparent irregularities in the orbits of Kuiper belt objects, so the hypothesis is not without foundation and may well turn out to be true. However, until someone actually finds this planet, we have no reason to believe or not to believe that such a planet exists (the irregularities that Planet Nine is supposed to account for have other possible explanations, cf., e.g. Shankman et al. 2017). Essentially, its existence is an article of faith, something that should clearly be avoided in science.1

Nevertheless, existential statements play a crucial role in scientific enquiry – note that we make existential statements every time we postulate and define a construct. As pointed out above, the statement in (1) rests, for example, on the statement "Words exist". This is an existential statement, whose precise content depends on how our model defines words. One frequently-proposed definition is that words are "the smallest units that can form an utterance on their own" (Matthews 2014: 436), so "Words exist" could be rephrased as "There is at least one x such that x can form an utterance on its own" (which assumes an additional existential statement defining utterance, and so on). In other words, scientific enquiry rests on a large number of existential statements that are themselves rarely questioned as long as they are useful in postulating meaningful hypotheses about our research objects.

Second, and more importantly, even if such confounding variables could be ruled out, no amount of data following the distribution in Table 3.2 could ever verify the hypotheses: no matter how many cases of windscreen we find in British but not American English and of windshield in American but not in British English, we can never conclude that the former cannot occur in American or the latter in British English. No matter how many observations we make, we cannot exclude the possibility that our next observation will be of the word windscreen in American English or of the word windshield in British English. This would be true even if we could somehow look at the entirety of British and American English at any given point in time, because new instances of the two varieties are being created all the time. In other words, we cannot verify the hypotheses in (2) and (3) at all. In contrast, we only have to find a single example of windshield in British or windscreen in American English to falsify them. Universal statements are a kind of mirror image of existential statements. We can verify the latter (in theory) by finding the entity whose existence we claim (such as Planet Nine in our solar system or a word for the forward-facing window of a car in English), but we cannot falsify them by not finding this entity. In contrast, we can falsify the former (in theory) by finding the intersection of values whose existence we deny (such as non-white swans or the word windscreen in American English), but we cannot verify them by finding intersections whose existence we affirm. Thus, to test a scientific hypothesis, we have to specify cases that should not exist if the hypothesis were true, and then do our best to find such cases. As Popper puts it: "Every 'good' scientific theory is a prohibition: it forbids certain things to happen", and "[e]very genuine test of a theory is an attempt to falsify it, or to refute it" (Popper 1963: 36). The harder we try to find such cases but fail to do so, the more certain we can be that our hypothesis is correct. But no matter how hard we look, we must learn to accept that we can never be absolutely certain: in science, a "fact" is simply a hypothesis that has not yet been falsified. This may seem disappointing, but science has made substantial advances despite (or perhaps because) scientists accept that there is no certainty when it comes to truth. In contrast, a single counterexample will give us the certainty that our hypothesis is false. Incidentally, our attempts to falsify a hypothesis will often turn up evidence that appears to confirm it – for example, the more data we search in an attempt to find examples of the word windshield in British English, the more cases of windscreen we will come across. It would be strange to disregard this confirming evidence, and even Popper does not ask us to: however, he insists that in order to count as confirming evidence (or "corroborating evidence", as he calls it), it must be the result of "a serious but unsuccessful attempt to falsify the theory" (Popper 1963: 36). In our example, we would have to take the largest corpora of British and American English we can find and search them for counterexamples to our hypothesis (i.e., the intersections marked by crosses in Table 3.2). As long as we do not find them (and as long as we find corroborating evidence in the process), we are justified in assuming a dialectal difference, but we are never justified in claiming to have proven such a difference. Incidentally, we do indeed find such counterexamples in this case if we increase our samples: The 100-million word British National Corpus contains 33 cases of the word windshield (as opposed to 451 cases of windscreen), though

some of them refer to forward-facing windows of aircraft rather than cars; conversely the 450-million-word Corpus of Current American English contains 205 cases of windscreen (as opposed to 2909 cases of windshield).

Reference

1. Anderson, N. (1999). Exploring second language reading. Boston: Heinle & Heinle.
2. Bransford, J. D., & Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. G. Chase (Ed.), Visual information processing (pp. 383–438).
3. New York: Academic Press. Day, R. (Ed.). (1993). New ways in teaching reading.
4. Alexandria, VA: TESOL. Day, R., & Bamford, J. (1998). Extensive reading in the second language classroom. New York: Cambridge University Press.
5. Smith, F. (1971). Understanding reading. New York: Holt, Rinehart & Winston.
6. Smith, F. (1975). Comprehension and learning. New York: Holt, Rinehart & Winston.