SQUIB

New York, Dayton (Ohio), and the Raw Frequency Fallacy*

ANATOL STEFANOWITSCH

There is a long-standing tradition in Chomskyan generative grammar of rejecting the relevance of corpus studies. A variety of arguments are put forth to justify this rejection, most importantly, that corpora are necessarily "finite and somewhat accidental" while the set of grammatical utterances is "presumably infinite" (Chomsky 1957: 15), and that, therefore, "probabilistic considerations have nothing to do with grammar" (Chomsky 1964[1962]: 215, n. 1; cf. also Chomsky 1957: 17). Chomsky is frequently reported as backing up this claim with the observation that

> the sentence *I live in New York* is fundamentally more likely than *I live in Dayton, Ohio* purely by virtue of the fact that there are more people likely to say the former than the latter (McEnery and Wilson 2001: 10).[1]

As always, it is difficult to decide whether Chomsky seriously offers this example in support of his position. Not that it really matters: Chomsky's contempt for − and his ignorance of − quantitative issues is of no concern to modern corpus linguistics. Chomsky's irredeemably anti-empirical views are firmly rooted in his anti-empiricist philosophy, and no amount of quantitatively sophisticated corpus-based argumentation will ever change his mind.

What is more troubling is the fact that many corpus linguists seem to be willing to accept Chomsky's quip as a valid argument about the skewedness of natural corpora (McEnery and Wilson 2001: 10), or the fallibility of frequency information (Fail 2004; McEnery and Wilson 2001: 10; Nelson 2002).[2] Instead of accepting Chomsky's argument, corpus linguists should refute it on at least two grounds.

First, and very obviously, corpus grammarians are not − and never have been − concerned with the frequency of individual sentences, but rather with the frequency of sentence *patterns*. This counterargument is

Corpus Linguistics and Linguistic Theory 1−2 (2005), 295−301 1613-7027/05/0001−0295

so self-evident that it needs no further comment. Second, and perhaps less obviously, Chomsky is misrepresenting the way in which corpus linguists make use of frequencies. Chomsky has identified what we might call the *observed-frequency fallacy*, which could be characterized as follows:

> *The observed-frequency fallacy*: Observed frequencies of occurrence represent relevant facts for scientific analysis.

I do not wish to deny that there are corpus linguists who are snared by this fallacy, but most of us reject it on the basis of what we could call the *expected-frequency epiphany*, an insight that can be characterized as follows:

> *The expected-frequency epiphany*: Observed frequencies of occurrence must be evaluated against their expected frequencies of occurrence before they become relevant facts for scientific analysis.

This insight is not particularly new or radical, nor does it apply exclusively to the field of corpus linguistics. Instead, it is the fundamental principle on which all inferential statistical procedures are based – procedures that are used routinely and as a matter of course in the cognitive sciences, the social sciences, the life sciences and even the physical sciences that Chomsky likes to quote as an ideal for linguistic inquiry.[3]

In the remainder of this squib, I will show that if we approach the sentences cited by Chomsky on the basis of the expected-frequency epiphany, the apparent problem that they illustrate disappears. Corpus linguists will no longer have to feel bad about the fact that the sentence *I live in New York* is, undeniably and irrefutably, more likely to occur – and thus more frequent – than the sentence *I live in Dayton, Ohio*.

In order to show this, we must first establish the observed frequencies of these two sentences. Clearly, we need a very large corpus in order to do this – the British National Corpus, with its one-hundred million words, contains a single example of the sentence *I live in New York* and no example of *I live in Dayton*, *Ohio* (perhaps not surprisingly, as it *is* a British corpus).[4] Presumably, the only corpus that is big enough for this kind of investigation is the Internet. In order to capture specifically American English usage, I restricted my investigation to the sub-domain of the Internet that uses the country suffix of the United States, <*.us*>.

To determine the relevant frequencies, I then proceeded as follows. For both sentences, I determined the total number of hits returned by two major search engines, Google and Alltheweb. In each case, I chose the higher of the two values. The rationale behind this procedure is that

a search engine may miss potential hits, but it should not be able to find more hits than are actually there; thus, the higher value should always be closer to the actual one. Of course, it would not have been plausible to simply search for the strings [I live in New York] and [I live in Dayton, Ohio]. There were three problems to be taken into account. The first problem is that [New York] may refer to 'New York City' or to 'New York State', and it forms the first part of both strings. Thus, searching for the string [I live in New York] would turn up hits for both [I live in New York City] *and* [I live in New York State], and the hits that contain neither the string [City] nor the string [State] would be ambiguous between the two readings. Since *Dayton, Ohio* is a city, I assumed that the comparison should be made with *New York City* rather than *New York State*. I solved this problem by including only hits that unambiguously referred to New York City, namely the strings [New York City] and [City of New York]. The second problem is relatively trivial: there are orthographic variants that must be included. For New York City, I additionally included [N.Y.C.] and [NYC], for Dayton I included [Dayton, Ohio] and [Dayton, OH]. The third problem was that although city names are typically followed by state names or suffixes in American English usage, this addition can be omitted when it is obvious which city is being talked about. This means that some cases of the string [Dayton] refer to Dayton, Ohio even though they are not followed by the strings [Ohio] or [OH], while the majority presumably refers to other cities called *Dayton* (there are, unfortunately, many such cities in the U.S., for example, in Minnesota, Kentucky, Virginia and Texas). It would seem that the addition *Oh(io)* will most likely be omitted from the name *Dayton* when one Ohioan is talking to another. I therefore solved this problem by including all cases of the string [Dayton] occurring on web pages with the Ohio state suffix <*.oh.us*> in addition to the strings [Dayton, Ohio] and [Dayton, OH] from all other <*.us*> pages. The results of this procedure are shown in Table 1 (in the following, I will use uppercase NEW YORK and DAYTON to refer to the respective set of strings).

Table 1. *Observed frequencies (Corpus-based)*

	Frequency
I live in NEW YORK (= … New York City, … the City of New York, … N.Y.C., … NYC)	566
I live in DAYTON (= … Dayton, OH, … Dayton, Ohio, … Dayton [site:.oh.uk])	12

These results show what we would have expected: *I live in NEW YORK* is indeed vastly more frequent than *I live in DAYTON*. But do these frequencies differ from the expected ones, i. e. could a corpus linguist potentially be fooled into attaching linguistic importance to them?

There are two ways in which we can determine the expected frequencies: first, on the basis of demographical information, i. e. the population sizes of the two cities, as suggested implicitly in Chomsky (1962[1964]) and explicitly by McEnery and Wilson (2001: 10); or, second, on the basis of corpus-internal information, i. e. using the base frequencies of the strings and string sets [I live in], [NEW YORK] and [DAYTON]. The first possibility is somewhat unusual in the context of corpus-linguistic research, as the relevant information about the frequency of objects in the external world is not usually available; the second possibility, in contrast, is standard procedure in corpus linguistics.

Let us therefore begin with the second option. The base frequencies for NEW YORK and DAYTON were determined using the same procedure and the same sets of strings as above. Table 2 shows all relevant frequencies (those in italics were derived from the web search, the others are the result of additions and subtractions). In addition, the expected frequencies for each cell are shown in parentheses – these were derived by the standard procedure of multiplying the marginal frequencies for each cell and dividing the result by the table total.

Table 2. *Observed and expected frequencies (Corpus-based)*

	NEW YORK		DAYTON		Total
I live in …	*566*	(563)	*12*	(15)	578
¬ *I live in …*	5,979,434	(5,979,437)	163,988	(163,985)	6,143,422
Total	*5,980,000*		*164,000*		6,144,000

Clearly, the observed frequencies of the two sentences differ only minimally from the expected ones, and the differences are not statistically significant (Fisher–Yates Exact Test, $p = 0.23$, n.s.).

Next, let us turn to the second option for statistical evaluation. In order to derive the expected frequency of the sentences *I live in NEW YORK* and *I live in DAYTON* from the population sizes of the two cities, we need reliable, up-to-date demographical information. Such information is provided by the *State and County Characteristic Population Estimates* (U.S. Census Bureau 2004). Table 3 shows the relevant information. The expected frequencies were derived as follows: the joint population size of the two cities is 8,264,372, where the population of New York City accounts for 98.06 percent and that of Dayton, Ohio for 1.94 percent. The joint frequency of the sentences *I live in NEW YORK* and *I live in DAYTON* is 578. Applying the proportions of the population sizes to the sentences, we get the expected frequencies shown.

Table 3. *Observed and expected frequencies (Population-based)*

	Population		*I live in …*	
	Absolute	Relative	Obs.	Exp.
NEW YORK	8,104,079	(98.06%)	566	567
DAYTON	160,293	(1.94%)	12	11
Total	8,264,372	(100.00%)	578	

Again, the differences between the observed and the expected frequencies are minimal, and again, they are not statistically significant (Binomial Test, $p = 0.45$, n.s.). In fact, the precision with which the population sizes of the two cities predict the frequency of the two sentences is somewhat eerie − I, for one, have never encountered such accurate predictions in the field of corpus linguistics.

The results are thus quite unequivocal: compared to the sentence *I live in Dayton, Ohio*, the sentence *I live in New York* is neither more frequent than expected on the basis of the frequency of the component parts [I live in], [NEW YORK] and [DAYTON], nor is it more frequent than expected on the basis of the population sizes of New York City and Dayton, Ohio. In other words, no linguistic theory, corpus-based or not, will have to worry about these two sentences and their diverging raw frequencies.

However, from these results it does not follow in any way that the 'importance of probabilistic considerations' has been or is being overrated or that natural corpora are so 'finite and accidental' as to be useless for the purposes of linguistic analysis. To claim this seriously would show deliberate disregard for the amazing insights into linguistic structure that frequency-based analysis has yielded (one of my favorite examples is Krug 2003, Section 4, which shows that the frequency with which modals are followed by the morpheme NOT predicts their degree of coalescence with that morpheme, and which thus starts from an observation that is not unlike Chomsky's).

What *does* follow from the results reported here is that statistical evaluation should be a *sine qua non* in corpus linguistics − if we need Chomsky (of all people) to remind us of this, then so be it.

Received July 2005
Revisions received August 2005
Final acceptance August 2005

University of Bremen

Notes

* As always, thanks are due to Stefan Gries for discussion. I would like to stress that none of what follows is his fault. *Correspondence address*: <stefanowitsch@uni-bremen.de>.

1. I should point out that I was unable to confirm this quotation: McEnery and Wilson (2001) attribute it to a paper presented at the Third Texas Conference on Problems of Linguistic Analysis in English (Universtity of Texas, Austin, 1958) and Halliday (1991: 30) attributes it to a lecture at the Linguistic Society of America Summer Institute (University of Bloomington, July 1964); the published version of the Texas paper (Chomsky 1964[1962]) contains a slightly different quotation:

   > [S]urely it is not a matter of concern for the grammar of English that "New York" is more probable than "Nevada" in the context "I come from −." In general, the importance of probabilistic considerations seems to me to have been highly overrated in recent discussions of linguistic theory. (Chomsky 1964 [1962]: 215, n. 10).

   It seems likely, then, that Chomsky used the *Dayton, Ohio* example in several lectures during the late fifties and early sixties and only switched to *Nevada* after the example had entered corpus-linguistic folklore. I will stick with the *Dayton, Ohio* version here, since this is the one that is invariably cited in the literature (the only exception I have come across is Wasow [2002]).
2. Both arguments are also related to the apparent problem that corpora do not contain negative evidence (Chomsky 1957: 16−17, cf. also McEnery and Wilson 2001: 11−12); I will comment on this problem in Stefanowitsch (to appear).
3. It does not follow from the expected-frequency, incidentally, that observed frequencies are irrelevant − experience tells us that, at least with respect to language, they are highly relevant most of the time. The problem is that, without statistical evaluation, we cannot distinguish those cases that are relevant from those that are not.
4. For the sake of completeness: the BNC contains two other North American cities following the string [I live in] − San Francisco and Los Angeles. More to the point, it contains four examples of *I live in (the centre of/Greater) London* and only one example each for the much smaller cities *Liverpool*, *Oxford*, and *Manchester* (among others). Overall, it contains 147 instances of the string [I live in], 62 of which are followed by a location name. The most frequent word following this string, however, is *hope* (5 times).

Data sources

Alltheweb. Internet search engine, available online at <http://www.alltheweb.com>.
British National Corpus, World Edition. Oxford: BNC Consortium.
Google. Internet search engine, available online at <http://www.google.com>.

References

Chomsky, Noam
1957 *Syntactic Structures*. The Hague: Mouton.
1962 A transformational approach to syntax. In Hill, Archibald A. (ed.), *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis*

in English 1958, 124−58. Austin, TX: The University of Texas. [Reprinted in: Jerrold A. Fodor and Jerry J. Katz (eds.), 1964. *The Structure of Language*. Englewood Cliffs, NJ: Prentice-Hall, 211−241.]

Fail, Lia
2004 Corpus Linguistics (2): The Corpus Approach. Available online at <http://www.proz.com/howto/174>, last access: August 15, 2005.

Halliday, Michael A.K.
1991 Corpus studies and probabilistic grammar. In Karen Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. Burnt Mill/New York: Longman, 30− 43.

Krug, Manfred
2003 Frequency as a determinant in grammatical variation and change. In Günter Rohdenburg and Britta Mohndorf (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 1−67.

McEnery, Tony, and Andrew Wilson
2001 *Corpus Linguistics. An Introduction*. Second edition. Edinburgh: Edinburgh University Press.

Nelson, Michael
2002 Business English Students, Teachers and Material's Writers: A Corpus-Based Examination of Anticipated and Reported Need. Available online at <http://www.kielikanava.com/corpuscourse.htm>, last access: August 15, 2005.

U.S. Census Bureau
2004 State and County Characteristic Population Estimates. Available online at <http://www.census.gov/popest/estimates.php>, last access: August 15, 2005.

Stefanowitsch, Anatol
to appear Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2.1.

Wasow,Tom
2002 *Postverbal Behavior*. Stanford: CSLI.