

## The empirical trend: ten years on

Geoffrey Sampson

University of South Africa

Linguistics Department

PO Box 392

UNISA 0003

South Africa

sampson@cantab.net

### **Abstract**

Linguistic science of the past half-century has often been distorted through neglect of normal scientific standards of empirical falsifiability. An earlier paper in this *Journal* used a quantitative literature survey to examine how far in practice the newer trend towards use by linguists of corpora and other empirical data sources had progressed. The result was ambiguous: a trend towards greater empiricism had occurred since about 1970, but around the turn of the century it appeared to have reversed, and the end of the period surveyed (2002) fell so soon thereafter that it was hard to guess whether this reversal was a blip or a long-term change. With a further decade of linguistic literature to examine, the present paper repeats the survey using a more systematic sampling technique, and this yields results that are much more clearcut than those of the earlier paper.

## 1. Two kinds of empiricism

There are two facets to the scientific concept of empiricism. Both are relevant to linguistics, but the literature of our subject has not always been explicit about the distinction. Clearly, “empiricism” is always about laying emphasis on the role of experience of the external world in the growth of knowledge. (The word derives from Greek *empeiria*, experience.) But, on one hand, the principle holds that anyone, as a child beginning to learn while still too young to understand verbal teaching, *does* as a matter of fact gain knowledge mainly by distilling it from sense-data, rather than by extracting knowledge from mental mechanisms that are built-in at birth (perhaps coded in the child’s DNA). On the other hand, the empiricist principle urges that those whose social role is to generate new knowledge (e.g. scholars, scientists) *ought* to make their discourses accountable chiefly to interpersonally-observable data, rather than treating great names of the past as authoritative or relying on aprioristic argument and speculation.

These two aspects of empiricism, which we might respectively call *descriptive empiricism* and *normative empiricism*, are clearly interrelated; but they are in principle quite separable. For instance, whether or not it is likely in practice, logically it is entirely conceivable that psychologists, studying the behaviour of infants using standard scientific techniques of observation, hypothesis-testing, etc., might discover that children gain significant swathes of factual knowledge without any exposure to relevant experience (or teaching), purely as a consequence of the operation of innate cognitive equipment. If that were to happen, empiricism in the normative sense would have undermined descriptive empiricism.

Rejection of descriptive empiricism is called either *rationalism* or *nativism*. (The logic of these names is that knowledge stems not from experience but from the faculty of reason, Latin *ratio*, and that knowledge is “native”, i.e. inborn, in us rather than acquired after birth.) That view of human nature was associated historically with Plato and with Descartes; in our own time it has been promoted by Noam Chomsky.

## **2. Universal Grammar versus empiricism**

Until the 1960s, linguists took empiricism (in both senses) for granted. Infants were seen as acquiring their first language by listening to the speech of their elders and imitating it (descriptive empiricism), and linguists were expected to pursue their researches by observing speakers' behaviour and using it to test the validity of linguistic theories (normative empiricism). However, Chomsky disagreed with (both kinds of) empiricist assumptions, and succeeded for many years in carrying much of the discipline with him. Chomsky's anti-empirical linguistic stance is better called “nativist” than “rationalist”: he stresses the claim that much of language structure is innate in the human mind, without linking that claim to the particular mental faculty of reason. Chomsky's linguistic nativism was briefly adumbrated in his 1964 book *Current Issues in Linguistic Theory*, and was developed at length in later writings.

Chomsky's linguistics contradicted descriptive empiricism by claiming that the child's ability to master its mother tongue depends chiefly on much of the structure of language being encoded in our genetic inheritance, as our bodily structure uncontroversially is – this “innate knowledge of language” needs only the trigger of exposure to speech in order to be aroused and applied by the growing child,

and the only specific features of the mother tongue which have to be learned from experience are those details which differentiate one language from another.

(Chomsky, perhaps remarkably, held that these features are fairly trivial as compared to the properties which are common to all human languages.)

Chomsky also rejected normative empiricism as a recipe for doing linguistic research. There is obviously a difference between linguistics and sciences such as, say, meteorology, or marine biology: a student of the latter is considering things which lie entirely outside himself, but an English-speaking linguist describing some aspect of the English language is discussing, among other things, features of his own cognitive functioning. This is uncontroversial; but it led Chomsky to hold that grammatical description need not be based on objective evidence, because a native speaker can look inside his own mind and extract reliable “intuitions” about what he can and cannot say.

Chomsky’s rejection of descriptive empiricism was revived and reinforced a generation later by Steven Pinker in his book *The Language Instinct* (1994), which made linguistic nativism accessible to an audience far wider than Chomsky’s readership. Although both Chomsky and Pinker base their arguments for innate knowledge mainly on facts about language and language-acquisition, for both of them language is merely an aspect of cognition where the evidence for innateness happens to be specially clear. They believe that in other areas of cognition also, the contents of our minds are largely what our genetic endowment makes them.

This turning away from descriptive empiricism became noticeable in a variety of human sciences towards the end of the twentieth century, and was recognized (e.g. by Donald Broadbent, 1973: 189) as stemming from the influence of Chomsky’s

linguistics. (Many strands of late-twentieth-century academic discourse turned away from normative empiricism also, though there the influence of linguistics was probably not significant.) Maurice Gross (1979: 861) saw linguists' belief in rationalist psychology as a consequence of their rejection of normative empiricism:

G[enerative] G[rammar] could have been demonstrated to be a descriptive method far superior to all previous traditional and structural attempts. But the insistence on an experimental paradigm which depends entirely on introspection to provide the linguistic examples ... has caused the field to evolve toward some surprising philosophical speculations. Work based on sentences demonstrably acceptable to all but a few speakers ... has almost entirely vanished. Academic discussions on forms of Universal Grammar have appeared instead.

Gross made his point in general terms, but others have given concrete examples of how reliance on introspection has misled linguists. One very striking case was discussed by Pullum and Scholz (2002), who documented the way in which far-reaching claims by generative linguists about "tacit" grammatical knowledge being innate in the human species were based very heavily on a single alleged fact: to acquire the correct rule for forming English questions in the absence of innate knowledge, it was argued, a child would need to be exposed to a certain type of question which was claimed to be extremely rare in practice – so rare that innate knowledge offers the only reasonable explanation for children's success in mastering the question rule. Chomsky had asserted that "you can easily live your whole life

without ever producing a relevant example ... you can go over a vast amount of data of experience without ever finding such a case” (Piattelli-Palmarini 1980: 114–15); for him the belief that each child encounters relevant evidence “strains credulity” (Chomsky 1976: 213). And the same claim has been repeated by linguist after linguist quoted by Pullum and Scholz, over a period of decades. Yet none of these linguists have suggested that they had checked this rarity claim against observational data.

Pullum and Scholz, referring to corpus data, suggested that the relevant form of question might occur more frequently than generative linguists supposed. Using data from the “demographically-sampled speech” section of the British National Corpus, representing the casual, spontaneous speech of a cross-section of the British population, Sampson (2002: 81, and cf. 2005b: 81) calculated that an average rate at which one could expect to hear the forms at issue must be at least once every few days; the expected number of instances encountered in a lifetime would be in the thousands.

Apparently, a well-nigh revolutionary new theory about human psychology had been based in significant part on the fact that linguists were willing to rely on their intuitions for information about how speakers use their language, and those intuitions could be wildly wrong.

### **3. A trend interrupted?**

The recent rise in interest in corpus-based research methods has been caused in part at least by a reaction against that unempirical style of linguistic research. But linguists who would identify themselves as “corpus linguists” are still, surely, a

minority, and it is not clear to what extent the discipline as a whole has genuinely reformed and accepted the logic of empirical methodology (normative empiricism).

In Sampson (2005a) I attempted to answer that question via a statistical literature survey, using the journal *Language* (widely recognized as the world's leading general linguistics journal, and one which aims to avoid partisanship and publish the best research emanating from any school of linguistics). That survey covered the years up to 2002; it showed, in brief, that use of empirical methods had indeed fallen to a low in the years around 1970 and had risen significantly since then. However, the rise was not steady. Figure 3 of Sampson (2005a) showed a bar chart which rose to a high in the mid-1990s and later fell back somewhat. And indeed, quotations from the linguistics literature of the time suggested that influential believers in the superiority of introspection over empirical observation as a research technique were challenging the trend towards empirical work. For instance, reviewing a corpus-based study by Rosamund Moon, Thomas Nunnally (2002: 177) wrote:

it is intuition that signals ill-formedness, not frequency of formations per million words ... [Moon's] shortchanging of native speaker understanding as evidence until a massive corpus can locate, say, five examples is worrying.

In 2003 Frederick Newmeyer even used his presidential address to the Linguistic Society of America in order to deprecate "usage-based" linguistics (Newmeyer 2003: 696). These were the first occasions I had encountered when mainstream generative

linguists attacked the corpus-based approach to research as explicitly undesirable, rather than merely ignoring it.

Indeed, one way of interpreting Figure 3 of the 2005a paper would be to say that, although the all-time low level of empirical work around 1970 was not maintained, after that level had risen gently to 1980 it subsequently remained essentially flat except for a blip (which might have been a chance matter) about 1995, and (except for that blip) it never returned to the level of 1950. (The year 1950 was sampled to give a baseline well before the issue of speaker's intuition as a valid data source had become part of the discourse of the discipline.) The survey had been carried out so soon after 1995 that it was not possible to discover whether a longer-term trend existed.

By now, a further decade of data is available, so that it is worth looking to see whether including the following ten years in the analysis might change the picture. And indeed it does, in a way that makes the picture much clearer than before. Without reopening the methodological issues discussed in Sampson (2005a), the present paper aims to show what has happened to the figures, when they are brought up to date.

#### **4. The new survey**

For the new survey I sampled *Language* volumes in the same fashion as before, but bringing the sampling up to the present, and regularizing the sampling basis so as to take two in every five volumes throughout the period since 1960. (In Sampson (2005a) the proportion of volumes sampled was higher after 1990 than for 1960–90, for reasons which seemed good at the time but which created a discontinuity that



potentially interfered with assessment of long-term trends.<sup>1)</sup> Sampson (2005a) discussed at some length the basis for categorizing a paper as empirically-based or not; it would be wasteful to repeat the details of this discussion here, but suffice it to say that the same guidelines were followed for the new sample volumes. In particular, the threshold defined in the original survey of two separate empirical data items required to count a paper as empirically-based was retained. That threshold is of course arbitrary, but so would any other threshold be, and for present purposes what matters is that it was applied consistently and in the same way as for the 2005a paper.

To repeat a point made in the earlier paper, “empirically-based” is not synonymous with “corpus-based”. Various kinds of interpersonally-observable data can potentially be relevant to different genres of linguistic research, and the interesting issue is how far linguistics is moving from introspection towards the use of interpersonally-observable data in general. Having said that, the recent increase in empirically-based research is in practice very largely accounted for by use of corpora.

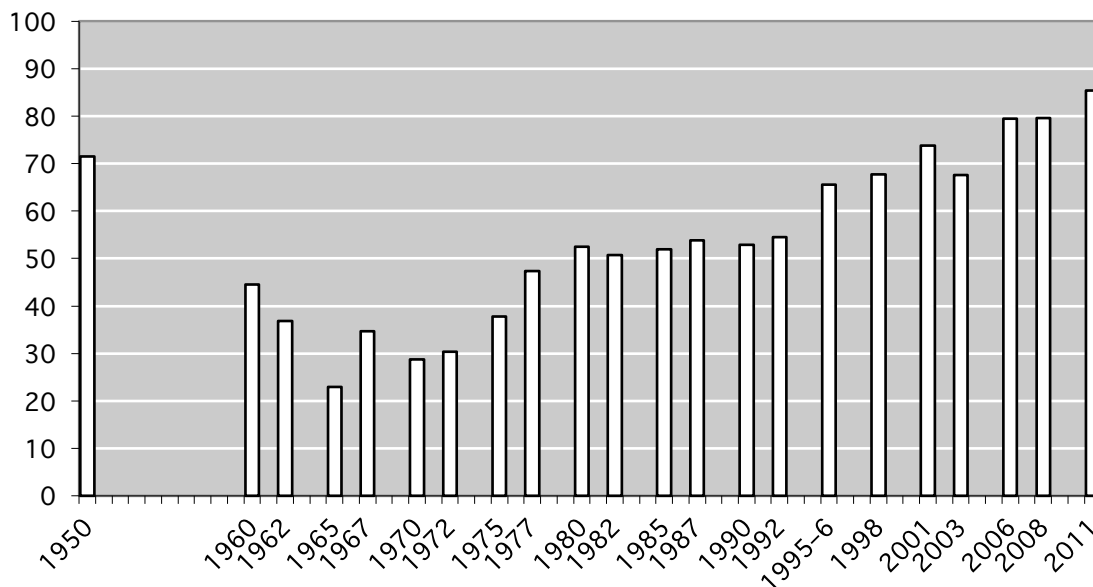


FIGURE 1 ABOUT HERE

Label axes as follows: x-axis “year”; y-axis “%”.

---

Figure 1 displays the new survey findings graphically in a manner directly comparable with Figure 3 of the 2005a paper. That is, the bar for each volume gives the number of empirically-based papers in that volume as a proportion of all empirically- or introspection-based papers, ignoring any “neutral” papers to which this distinction in research techniques is not relevant. (Sampson 2005a: 23–25 argued that many papers on phonological or morphological topics, or on the history of linguistics, among others, should be counted as “neutral”.) The figures for each

sample year between 1960 and 2011 are averaged with those of the preceding and following sample years (and the figures for 1960 and 2011 themselves are averaged with the sole adjacent year half-weighted), as a recognized statistical technique for revealing underlying trends in irregular time-series. (This averaging means that since 1990, when the sampling pattern changed to be more regular for the new survey, even bars for the same year in the two Figures will not necessarily be identical in length.<sup>2</sup>)

## 5. The blip washes out

Figure 1 speaks for itself. The 1990s “blip” and subsequent downturn in the 2005a paper have disappeared. What we now see is an increase, as steady as is plausible for such a complicated discipline on the borders of humanities and social sciences, from a low of empiricism around 1970 to levels which in recent years have comfortably exceeded that of the 1950 baseline. If it is appropriate to discuss the issue as a struggle between opposed methodologies, then it is a struggle which the empiricists are evidently winning. And this has not gone unnoticed by observers of the field. For instance in 2008 Brian Joseph, then editor of *Language*, remarked:

research papers are more experimentally based now than ever before ... Also, they are more corpus-based ... referees increasingly are demanding such multiple sources of data from authors who fail to offer them ... this trend towards “cleaner” and more “natural(istic)” data is all to the good of the field. (Joseph 2008: 687)

Notice that this very quotation argues against a sceptical objection that basing a survey of this kind on the contents of one journal might merely demonstrate partisan biases of that particular journal. I used *Language* alone because, with limited time available, that seemed to be the only practical way forward. But it is clear from the editor's comments that the increase in empiricism in the material he printed reflected preferences of the authors and of the wide range of referees used by *Language*, rather than being imposed by editorial prejudice.

## **6. Conclusion**

It seems, then, that (always assuming that the journal *Language* is adequately representative – and I believe it is) the unscientific trend which allowed numerous linguists to believe that language worked the way they wanted it to work is now rather definitely a thing of the past, among academic professionals.

That last qualification is important. Changes of mind on the part of an academic discipline take time to filter out to the wider audience of educated general readers (or even to members of neighbouring disciplines), so I believe that plenty of members of the public who take an interest in ideas about language are still heavily influenced by unempirical approaches and their “findings”. In this case the general time-lag which applies to all new intellectual trends is reinforced by a powerful special factor: speculative, intuition-based theorizing makes for easier reading than discussion based on detailed empirical data. The latter will inevitably involve numerous complications and qualifications which matter, but which are tedious for non-specialists to wade through (whereas linguists who rely on intuition are free to paint their pictures in broad and simple strokes). Writings addressed by empirical

linguists to wider audiences can only do so much to shield readers from those complexities, if they aim to be true to the underlying realities. On student courses, where it is not over-cynical to suggest that many hearers are more interested in acquiring just enough knowledge to get their degree than in chasing the truth down twisting byways, the speculative style of linguistics certainly goes down far better than the empirical style. (Claude Hagège noted as long ago as 1976 how language-teachers who study some linguistics as part of their professional training find the genre of linguistics criticized here relatively accessible and hence appealing – Hagège 1976: 17 note 1; and in practice language-teachers form a chief avenue via which academic linguists' ideas about language disseminate into the wider world.)

So it is comprehensible if the general public understanding of language and its nature continues to be distorted by the unempirical style of research which came into vogue in the 1960s. But the implication of this paper is that things are not always going to be like that. A current of opinion that is created by a novel academic trend cannot survive indefinitely unless it receives ongoing support from the survival of that trend, and unempirical linguistics is not surviving. Max Planck was one of many who have pointed out that a new scientific approach does not win acceptance by convincing opponents, but by the fact that the opponents eventually die off and a new generation grows up familiar with it (Planck 1949: 33–4). At present, unempirical linguistics is still being written and read. But it seems safe to say that this is a temporary state of affairs.

No doubt for some time to come there will remain holdouts putting forward arguments for the superiority of introspection as a source of data for linguistic

research, as we saw with Nunnally and Newmeyer in section 3 above, but history is not on their side.

## Notes

<sup>1</sup> Up to 1992 the present survey took years ending in 0, 2, 5, and 7; some quarterly issues of the 1995 *Language* volume happened to be unavailable to me, so I took four issues from the two years 1995 and 1996, and thereafter took years ending in 1, 3, 6, and 8.

<sup>2</sup> There are also differences between the two charts with respect to years 1960, 1962, and 1965. Rechecking revealed a clerical error in the calculations for those years, now corrected in Figure 1; fortunately, the error had not affected the overall trend of the 2005a chart.

## References

- Broadbent, D.E. 1973. *In Defence of Empirical Psychology*. London: Methuen.
- Chomsky, A.N. 1964. *Current Issues in Linguistic Theory*. The Hague: Mouton.
- Chomsky, A.N. 1976. *Reflections on Language*. London: Temple Smith.
- Gross, M. 1979. "On the failure of generative grammar". *Language*, 55 (4), 859–885.
- Hagège, C. 1976. *La Grammaire générative: réflexions critiques*. Paris: PUF.
- Joseph, B.D. 2008. "Last scene of all ..." *Language*, 84 (4), 686–690.
- Newmeyer, F.J. 2003. "Grammar is grammar and usage is usage". *Language*, 79 (4) 682–707.
- Nunnally, T.E. 2002. Review of Moon, *Fixed Expressions and Idioms in English*. *Language* 78 (1), 172–177.

- Piattelli-Palmarini, M. (Ed.) 1980. *Language and Learning: the debate between Jean Piaget and Noam Chomsky*. London: Routledge.
- Pinker, S. 1994. *The Language Instinct: the new science of language and mind*. New York: William Morrow.
- Planck, M. 1949. *Scientific Autobiography and Other Papers*. New York: Philosophical Library.
- Pullum, G.K. & B.C. Scholz. 2002. "Empirical assessment of stimulus proof arguments". In Ritter (2002), 9–50.
- Ritter, N.A. (Ed.) 2002. *A Review of the Poverty of Stimulus Argument*. A special issue of *The Linguistic Review*, 19 (1–2).
- Sampson, G.R. 2002. "Exploring the richness of the stimulus". In Ritter (2002), 73–104.
- Sampson, G.R. 2005a. "Quantifying the shift towards empirical methods". *International Journal of Corpus Linguistics*, 10 (1), 15–36.
- Sampson, G.R. 2005b. *The "Language Instinct" Debate*. London and New York: Continuum.