SHERLOCK HOLMES AND LEXICAL STATISTICS.

A late story by Dr Watson throws light on the origins of corpus linguistics.

One summer morning I had taken the early train from Paddington to a little town on the South Downs to visit my friend Sherlock Holmes. Since retiring from his life as a consulting detective, he had lived almost as a hermit in a farmhouse upon the moor above the town, withdrawn – as I had thought – from a life devoted to intellectual puzzles, and devoted entirely to the practical business of bee-keeping. It was but a short walk from the railway station to the farm, whose location was easy to discover. Even from afar it was signalled by wreaths of foul blue smoke curling forth high into the air from Holmes' old brier-root pipe.

As I arrived, he was sitting with his back to me, and I was confident that he had not noticed my silent approach, but he called out:

> "Watson, I note that you still need exercise, but that you have lost weight."

I asked how he could possibly know that it was I, and that I had lost weight.

> "Of course, I have the advantage of knowing your habits. Nevertheless, elementary! As to the first:
>
> (a) I see by the way my bees are swarming, that someone is approaching.
> (b) I conclude that it is you, since it is precisely 172 minutes since the departure of your preferred 7:15 train from Paddington. That makes 152 minutes for the train journey plus 20 minutes for the walk from the station.
>
> "As to the second:
>
> (a) I can hear that you are out of breath after your short walk, which should not require more than 15 minutes.
> (b) But I note that you passed silently through the narrow garden gate. The last time you called you almost wrenched the gate from its hinges trying to squeeze through.
>
> "Given these combined observations, I am confident that my conclusions are reliable. I would need a small piece of paper to calculate the exact confidence level, but I would place it somewhere around $p < 0.001$."

The curious phraseology in this last remark was new to me, but I let it pass without comment.

We sat outside in the warm morning sun, and whilst I idly swatted at the many bees which buzzed around, Holmes mused about aspects of the life of the mind which had not hitherto found a place in my little stories about his adventures. I had, of course, emphasized his contributions to logical analysis, and he had often taken a malicious pleasure in chastising me for muddling deduction, induction and abduction. I had also, in describing his investigation in Baskerville Hall, recorded his celebrated statement about scientific method. He rejected, if you recall, preposterous surmises and conjectures about supernatural dogs with glowing eyes which roamed upon the moor (the dogs, that is, not their eyes), and stated:

> "If we are dealing with forces outside the ordinary laws of Nature, there is an end of our investigation. But we are bound to exhaust all other hypotheses before falling back upon this one."

This clear formulation of what is known as "methodological naturalism", namely that explanations must avoid recourse to the supernatural and be limited to what can be observed and tested, had long become a universally accepted pillar of the philosophy of science. As he later remarked to me:

> "I am confident, Watson, that you will recognize this as a methodological maxim, and not confuse it with dogmatic metaphysical naturalism, which is an ontological claim about the non-existence of a non-physical world."

I assured my friend that it had not occurred to me to make this confusion.

> "I am confident that it hadn't," said Holmes drily.

I could often identify irony in my friend's remarks, even if I could not always with confidence identify their precise target.

More than once during the years that I had lived with him in Baker Street, I had observed both the irony and the vanity which underlay my companion's quiet and didactic manner. On that summer day, he told me at some length how his contributions to the foundations of logic had been recognized by honorary degrees from renowned European and American universities – he was particularly proud of a letter of recommendation from a colleague he called Charlie Peirce – which he accepted on condition that the universities provided a year's supply of his favourite disgustingly acrid tobacco. (I admit that, on that summer day upon the moor, the foul blue smoke did keep most of the bees at bay.)

He reminded me that I had also recorded how his linguistic abilities had solved several crimes. His observations on lexical frequencies in Bradshaw's Railway Guide and Whitaker's Almanack had provided the key to decoding two secret messages. They had also attracted the attention of text analysts, and his (admittedly eccentric) theory that Cornish was akin to the Chaldean language, along with his monograph on dating medieval manuscripts, had naturally come to the attention of historical linguists.

However, he now explained to me, it was his wholly unique proposal – he was indeed rather vain on occasion – to combine text analysis with his own developments of probability theory which led to a noteworthy intellectual breakthrough. His principle that "when one has eliminated the impossible, whatever remains, however improbable, must be the truth" was well known. But his crucial formal work on probability was developed when, on his small farm upon the moor, he took up bee-keeping. As I watched his bees fly around just outside the rank clouds of blue smoke, he mused:

> "The individual bee is an insoluble puzzle. But the aggregate becomes a mathematical certainty. You can never foretell whether any one bee will sting you, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant."

I recalled that, on a previous visit to the South Downs, I had been struck by several ill-dressed and brooding creatures in the White Hart Tavern in the village, poring over papers covered in strange symbols. This did not surprise me unduly, for I knew that, even after his retirement, Holmes had maintained his contacts with unsavoury elements from the dark jungle of criminal London. I assumed that this was a group of ne'er-do-wells working out a tax dodge, but, he explained, they were historical philologists, quantitative stylisticians, computational lexicographers, linguistic statisticians, and the like, whom he invited from time to time to join an exclusive research group, and with whom he was working on the most fundamental linguistic puzzle of all: How many words must a linguist analyse before all the outstanding puzzles of linguistics are solved?

> "You may be flattered to know, Watson, that the data which the group has used to test our initial empirical hypotheses is the corpus comprising the collected stories of my adventures which you yourself have written. Something over half a million running words. Not that your texts are a representative sample of the English language in general, but you use words, and one word – in the aggregate – is much like another, especially given the repetitiveness and predictability of your phraseology. It is therefore useful for our purposes."

I was uncertain as to whether I should be flattered to hear my literary creations described as "texts" and "data", and uncertain as to whether predictability was desirable in a story-teller, but I felt somehow proud and happy that my work could be of use to science. Nevertheless, I objected:

> "But surely, if you wish to understand all words, then you should study all the words in a big dictionary? A large but not infinite number."

> "Ah! Somehow you always manage to confuse several things at once. On this occasion you confuse both types and tokens, and lemmas and word-forms! When I talk of 'How many words?', I mean of course running words, as they actually occur in texts. That number, though not infinite in a technical sense, has nevertheless no upper bound."

Holmes' phraseology had indeed taken on a curious tinge. I tried – but failed – to imagine words running after each other down the page pursued by llamas.

> "We scholars talk of the ways and means whereby there is – in theory – no limit to the number of words which man can utter: you of all people should know this. Yet there is – in practice – a severe limit on the predictable sequences which actually occur. In your case the predictability is very predictable."

Again, I felt that the reference to my unbounded but predictable loquacity was somehow unkindly ironic. My expression must have revealed my confusion, for Holmes continued:

> "Let me put it this way. An individual bee may fly in your direction and sting you. But the swarm will ... Oh, never mind! Are you aware that in the *Baskerville* story alone you use the word *moor* on 163 occasions, and the phrase *(up)on the moor* on 62 occasions?"

It seemed to me that, since much of the action indeed took place upon the great rolling moor, these were very reasonable words to use. But I admitted that I was not aware of these precise frequencies, and he explained to me the unbridgeable gap between conscious knowledge and unconscious behaviour. He went on:

> "Your linguistic habits, I note, remain exactly as we might predict. In the present text the phrase *(up)on the moor* is used six times in only 1,600 running words – once every 266 words on average – over twice as frequently as in the *Baskerville* story. Given the small size of the samples, this is probably within the bounds of normal random variation."

Here again was his curiously tinged phraseology, whose meaning I felt that I had not entirely grasped. But I could find neither irony nor flaw in his chain of logical sequences. Indeed I noted that the extraordinary powers of my friend included the ability to perform arithmetic computations on a flow of language which I had not yet committed to paper.
