John Sinclair
University of California Berkeley

Too much meaning? - a corpus-driven perspective on language theory

An early report from corpus research was that language in use is not very tidy. Even the carefully prepared corpora of printed texts showed a great amount of variation in the realisation of expressions. In contrast, the growth in popularity of language formalism at around the same time emphasised a very different picture for the representation of language in the mind.

As corpora grew larger it became possible to mount studies in which meaning was central; meaning had faded almost out of sight in the more rigorous kinds of language description, and remained as an unexamined primitive in most others. Semantics developed separately from structural description because of the fundamental architecture of most semantic theories, and in any case has little relevance to language in use.

Corpus researchers gradually realised that a comprehensive and formal description of the full range of data patterns that were found in a large corpus was quite beyond any single coherent grammar, and any attempt to make such a description risked being branded a pseudoprocedure. To avoid embarrassing gaffes, a number of other grammars ? called local grammars ?

would have to be written, so there would be alternative descriptions, alternative meanings, of the same passage, large or small. The skill and experience of users would in most cases interpret the text adequately, but the precise positioning on a continuum of meaning between or among the rival analyses was a matter for the individual user.

At the same time it was emerging from corpus research that multi-word units of meaning were far more widespread and far more important than had been suspected. The precise identification of these units lies outside the boundaries of conscious perception, and the regularity that could be distilled from many instances was not apparent until large corpora began to give up their secrets. Multi-word units - which are organised lexically and not grammatically (despite the surface conformity that is often observed) - began to impinge on territory which had hitherto been reserved for syntax. Again, here were stretches of language for which there were two competing descriptions that could be confronted with one another.

There are many other indications that we should be thinking of alternatives to the tidy model of language, even in the inner recesses of the mind.