A1

We can finally discard the dichotomies between 'langue' versus 'parole' or 'competence' versus 'performance' and reassess the 'relationship between the potential of language and its instantiation' (Francis and Sinclair 1994: 194). Here we must be wary lest we be criticized for equating the entire language with any one corpus, however large. A corpus can never offer a final report of all things the native speakers of the language can say in principle; but, as the corpus increases in size and degree of detail, it can offer a steadily closer approximation of many things the native speakers are likely to say in typical contexts. Undeniably, speakers can and do say many unlikely things, such as 'square-cut simplicity' back in [116] (how can you cut simplicity?) but typically with an intuitive sense that these are unlikely and are thus suitable for creative variations against the background of the more likely things. Such is a leading strategy in the discourse of consumerism and advertising, which regales us with an endless parade of new descriptions for high-priced commodities, even when these are ostensibly 'simple'.

A large corpus powerfully refutes the old anxieties about a 'heterogeneous mass of speech facts' rife with 'fragments and deviant expressions', as we saw expressed by linguists like Saussure and Chomsky. Large corpora display strikingly delicate dialectical balances between heterogeneity and homogeneity, or between diversity and uniformity. A bit paradoxically, large-corpus data manifest both fine-tuned regularity and fine-tuned creativity; indeed, speakers can be most effectively creative when they have a delicate sense of the normal or typical.

Still, the regularities are not just due to 'rules' of the types postulated in conventional linguistics at the level of the overall system. Instead, we encounter complex arrays of data displaying the selections and combinations performed by a huge population of discourse participants. The lexical regularities can be aptly designated with Firth's terms 'colligation' for a 'syntagmatic relation' and 'mutual expectancy' among 'elements' of 'grammatical' 'structure'; and 'collocation' for lexical items 'presented in the company they usually keep' (cf. Firth 1968: 186, 111, 182f, 106ff, 113). Determining when selections and combinations might qualify as colligations and collocations is scarcely feasible without large corpus data; intuition is not detailed or delicate enough, though it vitally help us interpret and evaluate the data in the corpus. Even so, we will always have some borderline cases where the evidence is insufficient at the current size of the corpus, e.g., whether 'classic simplicity' in [112-13] might be a collocation.

The creativity revealed by corpus data is even further outside the purview of conventional linguistics. The data show that creativity is a continual factor in discourse precisely because many regularities of a language are only decided on the plane of the actual discourse. Speakers typically perform an array of choices which is, as a whole, highly improbable or even unique in a statistical sense, e.g., to produce a combination like 'wonderful sets of blocklike simplicity, exquisitely lit' in sample [119], yet which is readily produced and comprehended by speakers of the language. So corpus data animate us to reinterpret the concept of statistical probability in language: although the probability of a whole array may indeed be very low in respect to the whole language — or at least to the whole corpus, which is all we would compute — some choices can make others significantly more or less probable (cf. Halliday 1991, 1992). For example,

the COBUILD data examined above show how some contextual cues such as 'politics' make the choice of 'instability' far more probable than 'fluctuation', whereas other cues such as 'currency' do just the opposite. 'Instability' has been chosen to be a cover term for any major social change which might disrupt the status quo of power and inequality, whereas 'fluctuation' designates the leeway for the rich to get even richer on the money markets without real work.

For a very large corpus, the raw frequency of an item is much less significant than collocability, its potential to be a collocate of other expressions, as contrasted with collocation to designate 'frequent co-occurrence' in an actual corpus (cf. Greenbaum 1974: 80). A sequence would be more creative when our intuitions based on collocability not fulfilled. For example, 'indeterminate' would be more 'collocative' with 'haze' [225], 'blank space' [217], and 'period' [215] than with 'Gallic blue' [223] and 'date' [216], since these two last logically seem rather well-defined; strictly speaking, nobody can be 'born on an indeterminate date' [216], because being born cannot extend over days, months, or years, though some doubt may arise later on about just when it was. Also, I would intuitively not expect 'jealousy' to be combined with 'complexity', but I can see where we might revise our expectations in view of 'indeterminacy' and 'irrationality' in sample [11]. Evidently, otherwise improbable combinations can be integrated on the plane of the discourse, and unfamiliarity does not hinder comprehension.

Paradoxically perhaps, collocability itself can be both precise and non-deterministic, due to the enormous range of possible combinations. Even if we compiled and interpreted all the positions of an item, our results could not be

equivalent to its total collocability for the same reason that no corpus, however large, can be equivalent to the entire language, as I remarked. Since collocability can evolve and change, a monitor corpus that is continually updated like the COBUILD 'Bank of English' can help us keep track, e.g., when governments and banks have established the term 'fluctuation band', a collocation I would not have predicted from my intuition. By showing how the place of an item within the language system is non-deterministic and evolutionary, corpus research also shows why no set of deterministic 'semantic features' or similar constructs could definitively represent the whole system of 'possible meanings' in a natural language, nor, strictly speaking, even the 'meaning' of a single word. If meanings are always evolving, none can be impossible. At most, special and technical cases like 'deterministic nonperiodic flow' in the work of Edward N. Lorenz and others, which occurred 3 times in my COBUILD data, could attain a high stability and probability; but such cases are not representative and cannot promote coverage and consensus in a semantics of real language.