'If I were you...': Language Standards and Corpus Data in EFL

Robert de Beaugrande
Universidade Federal de Minas Gerais

Este artigo discute questões sobre o inglês padrão e o uso de corpora no contexto do ensino de inglês como língua estrangeira. Reflete-se sobre o papel que os textos autênticos passam a ter tanto para a lingüística quanto a lingüística aplicada. O autor ressalta que professores e alunos de inglês como língua estrangeira não podem esperar por um futuro no qual todos os problemas teóricos e práticos tenham sido resolvidos e convida todos para participar de forma ativa na construção de um ensino de línguas baseado em textos autênticos. Nesse sentido, ressalta que o input e feedback de todos os envolvidos no ensino de línguas em seus diversos estágios é crucial para que problemas sejam avaliados e resolvidos.

Standart English, authentic English

Periodically, 'standard English' becomes a topic of lively controversy that staunchly resists being reconciled. To appreciate why, we might identify two incompatible conceptions. In the *exclusive* conception, standard English is a pure medium with stringent rules and precise boundaries; all non-standard usages and variations are uniformly classified as *errors*. The 'standards' are perceived to be undergoing a *decline* unless strong measures are taken.

The exclusive conception encourages TESOL and TEFL to emphasise a strict division between 'standard' versus 'non-standard' English. In practice, teaching and learning are typically arranged to minimise the occurrences of non-standard English even when the learners' fluency is insufficient to produce standard English on their own initiative. So the initiative is denied by enlisting them in imitative production, such as repeating or writing down the teachers' utterances, or reading out samples of English from textbooks or worksheets. The samples sustain their 'standardness' by being patently simplified and uniform, like these I found in real textbooks:

[1] I hear with my ears.
[2] The red rose is fine.

[3] Useful knowledge is desired.
[4] I like a bedroom with green walls.

Insofar as this variety of standard English is not found in the everyday language use of native speakers, I propose to label it *non-authentic English*, and to question whether exposure to it, however prolonged, can be realistically expected to build fluency for the *authentic English* used by native speakers.

In the *inclusive* conception, Standard English is one variety bordering upon and overlapping with a family of 'World Englishes'. Any non-standard usage which widely occurs in some variety is not an error but a *variation,* and should be understood from its social, historical, or geographical motivations. The 'standards' are perceived to be undergoing not a decline but *diversification*, due to swift and massive increases in the size and distribution of the world-wide population of prospective learners of English. If a decline does ensue, then chiefly because this population cannot gain sufficient exposure to standard English that is also authentic English.

The inclusive conception encourages TESOL and TEFL to emphasise the family resemblances among varieties of English. One bundle of these varieties comprises 'learner Englishes', which, though partially non-standard, are nonetheless systemic. Their errors are natural variations within a transitional hybridised system where control is shared between the system of English and the system of the native language. Here are some UAEU fourth-year student data from self-paced writing assignments, which I have observed to be typical:

[5] The UAE desert rich with green plants [= The UAE desert is rich with green plants]
[6] Most of people in the world speaking English [= Most people in the world speak English]
[7] Teacher may be dirty from head to toes since the chalk. [= A teacher may be dusty from head to foot/from top to toe because of the chalk.]
[8] Being a politician is appropriate for women because they like argument and debate and can talk hours. [= talk for hours]

These students are predictably compensating for their limited fluency in authentic English by extrapolating from their native Arabic.

The latter language lacks 'be' in the Present [5];[1] *Webster's Seventh* uses Participles in place of Finite Verbs on many occasions [6]; lacks the Indefinite Article [7]; and uses Nouns in the Accusative Case as Adverbs of Time [8]. In the collocation 'head to toes' [7], the student-produced variation is more logical than Standard English, which says 'head to foot' (why not 'feet'?) or 'top to toe' (why not 'bottom'?) (*COBUILD* 670, 1544).[2]

The inclusive conception defines the long-range task of TESOL and TEFL as promoting a convergence between multiple systems: between learner Englishes and standard English. The convergence must be gradual because of the sheer size and complexity of the task; but these factors should seem less daunting if we have maps of the intermediary stages to identify signals of progress.

The exclusive conception, in contrast, can provide no maps of progress insofar as it views these stages as arbitrary fluctuations in the levels of error production. This view totally overlooks the systemic quality of learner Englishes, witness this magisterial pronouncement: 'very few speakers limit their aberrancies to the widely shared features; each individual typically adds in his own speech a large and idiosyncratic collection of features' (Prator, 1968: 464). Such a view reflects a deplorable lack of knowledge and respect regarding learner Englishes, and does not hold up in the face of systematic documentation (as in Granger [Ed.], 1998).

The same view distracts us from appreciating the impact of non-authentic English on the learning process. The simplified, uniform English thought appropriate for non-natives readily strays over the border where non-authentic English becomes non-standard English too, as in these samples (again from real textbooks):

[9] I differentiated milk from water.
[10] The plane fell but the loss is small.

---

[1] More precisely, 'Past' and 'Present' are misleading translations by Westerners for the Arabic completive and non-completive Aspects, which Arabic grammarians call 'maaDi' and 'muDaari'. This factor too causes my students great confusion. Also, the Enumerators like numbers and 'most' are mainly Nouns followed by the Genitive, whence the logic of 'most of people'.

[2] To conserve space, the wordy titles of dictionaries are shortened as follows:
COBUILD = *Collins COBUILD English Language Dictionary* (1987)
Random Webster = *Random House Webster's College Dictionary* (1991)
Webster's Seventh = *Webster's Seventh New Collegiate Dictionary* (1963)

[11] Be polite so that you could be acceptable.
[12] Would that world peace is permanent.

The authors were non-native speakers, but then so are many teachers of EFL, some of whom, finding these samples in a textbook, might let them pass. My point is that an easy acceptance of non-authentic English can dull our sensitivity for standards of English that are far more subtle than the routine issues in pronunciation, orthography, and grammar.

I would make the same point on a higher level when non-authentic English is used in theoretical courses in linguistics, such as syntax and semantics. There, the motive of the textbook authors is different: the linguistic analysis is so complicated and artificial as to be practicable at all only for simplified samples of English like the evergreen 'the man hit the ball'. But the outcome is much the same: learners are exposed to sentences which do nothing to enhance their fluency in authentic English. One textbook I encountered for a fourth-year university course in semantics demonstrated the conceptions of 'analytic', 'synthetic' and 'contradictory sentences' with data including these:

[13] John's nine-year old brother is a boy.
[14] The fly was on the wall, so the wall was under the fly.
[15] John is taller than himself.
[16] That girl is her own mother's mother.

Here, the samples are non-authentic because semantics proposes to study meaning whilst 'deliberately excluding any influence of context or situation of utterance' (Hurford and Heasley, 1996: 91). Preference goes to samples that neatly divide up between obviously true and obviously false. But the division is irrelevant to everyday conversation and also to the tasks of TEFL. We would justly feel absurd animating our students to go about uttering things like [13-14] and not uttering things like [15-16].

Authenticity is a problem whose subtlety and importance have yet to be fully recognised. Samples which reflect the systemic qualities of learner English would be authentic in terms of that system, as different from the isolated and truly 'idiosyncratic' occurrences when learners panic or make wild, random guesses. But the difference may not be easy to recognise unless we have large and systematic samplings of learners Englishes (see now Granger [Ed.], 1998). Moreover, our students produce many samples which, though not manifesting any errors or violating any

'rules', are not authentic standard English, e.g. (more UAEU student data):

[17] I am about to hate my major [= coming to hate my major]
[18] I will succeed if luck is present [= if I am lucky/if luck is on my side]
[19] I ask God to make me achieve my ambition [= I pray that God may grant me the achievement...]
[20] And we as women, our message is to rear our children excellent rearing. [= And for us women, the message is to rear our children excellently.]

These subtler problems again imply a tension among the 'standards' of English and the standards of the home language. Sample [12] reproduces a construction considered elegant in Classical Arabic, namely the Accusative Absolute, where a Verb takes a redundant Participial Object formed from the same stem ('rear a rearing'). Standard in Arabic too is making one Noun Phrase at the start of a sentence be the Topic and another be the grammatical Subject, also found in [12]. What appears to be careless errors is in fact the result of careful attention to transposed standards.

Theory and practice of large corpora in EFL

By my line of argument so far, learners of EFL, and some non-native teachers of EFL too, suffer not from exposure to non-standard English, but partly from exposure to non-authentic English and partly from lack of exposure to authentic standard English. Similarly, the real danger is not that the standards of English are in a decline but that those standards may be to a large extent unrecognised and inaccessible. The world-wide population of non-native learners cannot encounter enough authentic standard English to gain the intuitive control over its standards that native children achieve. Standard English is represented by a uniform and simplified variety of non-authentic English; standards are exclusively defined in terms of hard and fast 'rules' for every occasion. Under these conditions, the chances of genuine success heavily favour learners who have extensive outside exposure to authentic English, e.g., through satellite television or personal computers with Internet access.

I am using the term 'standards' here in a programmatically inclusive sense which I hold to be justified and realistic to the degree that these standards are documented by large sets of authentic English and not just asserted out of personal attitudes about correctness or propriety. They are sustained by the preferences of fluent speakers or writers for certain arrays of lexical and grammatical choices from among the immensely larger set of theoretically possible choices These include not just the choices stipulated by 'grammar' and 'vocabulary' in their routine senses, but also stylistic and rhetorical choices, as well as choices stipulated by genre, register, and text types. These standards *co-ordinate sets of choices*, so that what is chosen at one point makes certain choices at other points more probable. Only the more obvious and regular standards are reflected in textbooks relying on non-authentic English illustrated above, such as the major patterns of the English Noun Phrase. So, learner performance tends to feature non-authentic English too, which, however dull, seems safe.

Where the standards are comprehensively reflected is in suitably searched and sorted attestations of authentic English in large corpora. Their potential for language teaching is thus steadily gaining recognition, witness some recent collections and edited volumes (e.g. Botley, Glass, McEnery, and Wilson, 1996; Wichmann, Fligelstone, McEnery, and Knowles, 1997; Burnard and McEnery, 2000; Lewandowska-Tomaszczyk, and Melia, 2000; Ghadessy, Henry, and Rosebery, 2001). From a practical standpoint, the 'convergence' envisioned by Leech (1997) between 'language teaching' and 'language corpora' was probably inevitable in view of the manifest practical value of corpora for 'data-driven learning' (Willis, 1993; Johns, 1994). The value is plainest in areas of study where we are compelled to work with a great deal of data, such as style (Jackson, 1997), register (Biber, 1994), genre (Carne, 1996), and of course literature (Kowit and Carroll, 1991; Louw, 1997).

From a theoretical standpoint, however, we are still far from a convergence in our thinking about how corpus data can or should transform both theory and practice of language teaching (cf. discussions in Aston, 1995; Barlow, 1995; Tognini Bonelli 1996). Corpus data have long been used in studies of language, but these were not designed for the goal of teaching non-natives. Such was true of corpus data for research in 19th-century philology on European dialects (e.g. Wencker, 1887-95) and in 20th-century linguistics on American varieties of English (e.g.

Kurath, 1949) and on Native American languages (e.g. Sapir, 1922). This work was eminently practice-driven; theory sporadically took shape in higher-level statements about language types, as in Sapir's (1921) ambitious 'classification' of languages.[3]

As some writers on corpus work have remarked (e.g. McEnery and Wilson, 1996), corpus studies underwent a period of eclipse in the 1960s and 1970s. 'Descriptive' methods were substantially displaced by 'generative' ones; and linguistics transformed its subject matter away from data sets in particular languages (plural, count noun) over to a single theory embodied in language (singular, mass noun) (cf. Beaugrande, 1991, 1998). As an integral step in this transformation, it was argued that 'attempting to state methods of analysis that an investigator might actually use, if he had the time', for 'constructing the grammar, given a corpus of utterances', must 'fail to provide answers to many important questions about the nature of linguistic structure' (Chomsky, 1957: 51ff).

For us, the key reservation to recall was that 'the corpus of observed utterances' 'obtained by the linguist in his field work' is '*finite* and somewhat *accidental*' (Chomsky, 1957: 15, my emphasis). This reservation was accepted at face value without noticing that it holds for every set of observations and every set of data in every science. Whatever a science or scientist has 'observed' must be 'finite'; and 'data' are, both by definition and by etymology, 'the given', and cannot be other than finite.

And linguistics is after all not a science of the infinite. If a language were an 'infinite set of sentences' (Chomsky, 1957: 13), then the act of uttering or comprehending a sentence would require infinite search times. Moreover, 'performance' would be related to 'competence' in purely accidental ways, just as, in the familiar parable, a roomful of chimpanzees with typewriters would, in infinite time, write the works of Shakespeare. (They would also type the British National Corpus and the COBUILD Bank of English.) Such fanciful quibbles and quiddities inhere in the mathematically proper meaning of the 'infinite', which supplies poetic

[3] Sapir (1921: 142f) produced a chart of 25 languages 'based on the nature of the concepts expressed' ('simple, complex'), the 'degree of fusion' ('fusional, isolating, agglutinative'), and the 'degree of synthesis' ('analytic, synthetic, polysynthetic'). The terms go back to 19th century mentalist studies of language types (e.g., Steinthal 1860), and have not been widely used in modern linguistics after Sapir.

and philosophical labyrinths for writers like Jorge Luis Borges but merit no place in the theory of language.

A language must rather be a *very large but always finite set of data*. This set can never all be observed any more than can the set of particle collisions in physics or the set of supernovas in astronomy. Nor can the whole set be consistently described by any single definition of 'sentence', which is a reliable unit only in reference to the clause structures of written language (Beaugrande, 1999). And least of all could such a set be represented by the non-authentic sentences of 'the man hit the ball' type preferred by generative linguists for their complicated analyses, as I noted in section 1.

Scientists who work with very large data sets must manage a *trade-off* between *breadth* (how much data a theory can describe) and *depth* (what degrees of detail and precision the description can achieve). Early corpus studies of familiar languages (e.g. on speech varieties) could aim at the sweeping breadth of Wencker's 'language atlas' or Kurath's 'word geography' because the structure of the language was in any case under control. But corpus studies of an unfamiliar languages, e.g. the Yana language of Northern California described by Sapir, had to concentrate on depth to work out 'fundamental elements' of the structure under control, and the breadth was correspondingly limited. Yet if this 'limitation-in-principle to classification and organization of data' from a 'corpus of observed speech' 'establishes' the 'inadequacy' of the description (Chomsky 1965: 15, 67), then 'adequacy' must have some odd meaning, as we will in fact see in a moment. I cannot understand.

Now, if a language were an infinite set, then its description would entail an infinite breadth that flattens out our depth to an infinite shallowness, and our description (completed in infinite time, by the way) would capture only infinitesimal detail and precision. In practice, generative linguistics evaded its own 'infinity' argument against corpus studies by 'assuming that the set of grammatical sentences is somehow given in advance' (Chomsky 1957: 18, 54, 85, 103). Breadth was hypothetical, built into the theory by focusing on the 'ideal speaker-hearer in a completely homogeneous speech-community, who knows its language perfectly'; and on 'language universals' 'stated only in general linguistic theory as part of the definition of the notion "human language"' (Chomsky, 1965: 4, 6, 117), Breadth in the sense I suggest did not seem to figure on the agenda; Chomsky's well-known *Aspects* was presented

with just 24 non-authentic English sentences (or 'transformations' of these).

Science also enlists technologies to cope with *accidents* in our data, most crucially at frontiers where we can't yet distinguish the accidents from the regularities. We scan the collisions in linear accelerators for evanescent particles at the frontiers of physics; we train our telescopes on invisible planets affecting nearby stars in detectable ways at the frontiers of astronomy; and we peruse our monitors or print-outs for units or patterns of a language attested in very large corpora at the frontiers of linguistics.

All across science, the more significant the potential for accidents, the greater the breath we should seek, and the more we deploy those technologies that increase breadth without seriously decreasing depth. We may thereby push down the significance of any particular accident (or set of accidents) by measuring its probability. Should the probability remain high, then we are dealing with a regularity that had been mistaken for an accident.

The return of large corpora of authentic language to the centre of language study thus impels us to reopen the whole discussion of theory and practice. In that spirit, Sinclair (1999a, 1999b) has recently probed the concepts of 'observational', 'descriptive', and 'explanatory adequacy' introduced by generative linguistics (e.g. Chomsky, 1964, 1965). The first two of these had been stoutly affirmed in early corpus linguistics, although not under such programmatic labels. Observation was an operation of recording data so as to sustain validity and rigour despite the lack of technology. Description was a thorough presentation of the observed facts, e.g., by drawing maps to locate the distinct dialect forms in the respective regions.

Explanation remained a separate and sporadic issue, since questions about why a language assumes a given form or evolution were rightly judged intractable. At most, linguists hoped that explanations would eventually arrive, viz.: 'back of the face of the history are powerful drifts that move language, like other social products, to balanced patterns'; 'perhaps psychologists of the future' will find 'the ultimate reasons' (Sapir, 1921: 122). The real theoretical significance of early corpus work adhered in the programmatic acknowledgement of the importance of languages or dialects that had hitherto been regarded as curiosities or degradations, much as learner Englishes are regarded in some quarters today.

The generative approach expressly disavowed observational adequacy by announcing that the 'observed use of language' 'surely cannot constitute the subject-matter of linguistics' (Chomsky, 1965: 4). So 'descriptive' and 'explanatory' adequacy' were consigned to goals requiring no observation, namely 'describing the intrinsic competence' and 'intuition' of the 'idealized native speaker' (Chomsky, 1965: 34, 26). Yet despite ambitious claims, intuition is a weak, opportunistic technique. Instead of actively observing data, one passively rates the plausibility of data one produces for the occasion. To judge by the samples in published studies, intuition heavily favours what I have labelled non-authentic data. And just as non-authentic samples cannot be an adequate basis for learning the language, neither can they be an adequate basis for describing the language nor for explaining 'language' in the abstract.

In the new corpus linguistics, observational adequacy is our operational front end, where we depend most crucially on our techniques and technology. Large corpora offer us such immense breadth that depth can be managed only by incorporating our techniques into our technology — harnessing the computer for the descriptive stage as well. The key question here is what mode and degree of depth to look for. We cannot simply re-open the programme of early corpus linguistics insofar as we do not share its goals of describing either unfamiliar languages or dialects of familiar languages. Still less can we embrace the programme of recent non-corpus linguistics in quest of 'linguistic universals', which are not even expected to fit large data sets, viz.: 'if some remarkable flash of insight were suddenly to yield the absolutely true theory of universal grammar', 'it would be at once "refuted" by innumerable [infinite!?] observations from a wide range of languages' (Chomsky, 1980: 2).

The programme of corpus linguistics might do well to shelve the principle of explanatory adequacy as long as the conception of 'explanation' eludes and operational definition. A more tractable principle at present would be *applicatory adequacy*: how far our work is found suitable and productive for relevant applications. So far, the most successful applications, which have in fact quickly become the industry standard, have been achieved in reference works, such as dictionaries of words, idioms, phrasal verbs, and so on. Selecting entries by their frequency of attestation renders the breath of coverage fully operational, and finally trims off gratuitous arcane or archaic expressions for ordinary meanings still found in conventional dictionaries, such as 'operose' for

'involving much labour', or 'monopsony' for a 'market condition with one buyer and a large number of sellers' (*Random Webster* 948, 877).[2] English has accumulated a peculiar mass of these terms, which are especially non-authentic in never being used in real-life conversation.

In parallel, depth become operational as more precise information on usage, such as noting which Verbs are frequently used only in the Active (e.g. 'elude') or only in the Passive (e.g. 'construe') (*COBUILD* 458, 302)[4]. We do not assert that using such Verbs the other way round counts as 'error', but that it is not expected. In my own 10-million-word corpus of British and American writers dating roughly between 1750 and 1920,[5] I found only 2 out of 74 uses of 'elude' in the Passive:

[21] they lessen the consumption; the collection is eluded; and the product to the treasury is not so great (Alexander Hamilton)
[22] My importunities would not now be eluded (Charles Brockton Brown)

But 21 out of 59 uses of 'construe' were found in the Active, not just in the current sense of 'interpret' [23] but also the senses of 'translate' [24] or 'interpret something into something else' [25] — neither of which I would use nowadays.

[23] This behaviour in her niece the good lady construed to be an absolute breach (Fielding)
[24] he recalled the shrewd northern face of the rector who had taught him to construe the *Metamorphoses* of Ovid in a courtly English (Joyce)
[25] She's an excitable, nervous person: she construed her dream into an apparition (Charlotte Brontë)

---

[4] This information does not appear for these same entries in the 1991 *Random House Webster's College Dictionary*, even though the latter is 'founded' on a 'large database' (435, 292, vii).

[5] Actually, this consists of several distinct corpora I am still in the process of organising: British literature, like Austin, Dickens, and Wilde (3 million words); British academics like Darwin, Bulwer-Lytton, and J.S. Mill (2 million words); American literature like Hawthorne, Mark Twain, and Willa Cather (3 million words); and distinguished Americans like Thomas Jefferson, Jane Addams, and W.E.B. DuBois (2 million words). The sizes of the corpora and the choices of texts (each of them complete) depend on what I can download from Internet websites; and their offerings are in turn limited to works in the public domain.

Such findings highlight the historical dimension of authenticity, the more so for EFL programmes that focus on literature, which I shall discuss further on.

Corpus-based reference works also offer an operational measure of applicatory adequacy for description in terms of suitability for a real audience, including non-native speakers. Compare these definitions:

[26] hydroponics: the growing of plants in nutrient solutions with or without an inert medium to provide mechanical support (*Webster's Seventh* 408)[6]

[27] hydroponics is a method of growing plants in water rather than in soil (*COBUILD* 714)

In conventional dictionaries, definitions have been authored by specialists in the field. [26] was evidently composed by a botanist — technically correct but accessible only to other specialists, who would understand how an 'inert medium' can 'provide mechanical support' to a crop of tomatoes (and who would know the meaning of 'hydroponics' anyway). [27] clears away the technicalities and explains the essentials for ordinary people.

To be sure, dictionaries represent the most thoroughly practical application of corpus studies. The theory is sparse and straightforward: a language can be represented by a subset of expressions whose occurrences in a very large corpus reach a specified cut-off point; that quantity of occurrences is sufficient to determine the meaning; and the definitions are to be illustrated with authentic data, which are 'examples of good practice' for 'speaking and writing the English of today' (*COBUILD* xv). The dictionary can impose authenticity without having to explain its nature nor defend it against theoretical or applied linguists who deal in non-authentic data.

Yet the theoretical implications of authenticity surely extend much further. We must decide whether corpus studies will be fitted to established descriptions and categories of linguistics; or whether the foundations of linguistics will have to be revised in light of corpus studies (Tognini Bonelli,

---

[6] The data were kindly provided via Internet by Stephen Bullon, Publishing Manager. The full corpus is not open to overseas access, but only a 50-million word chunk of it — one tenth of the total size. I picked 10% of the data from the full corpus, which may or may not be roughly equivalent.

1996; Sinclair 1999). As we know from the work on 'scientific revolutions' in the philosophy of science since Kuhn (1970), a theory is not displaced by observation alone but only by another theory which fits observation better, or which enables new and important observations (e.g. Kuhn, 1970). Now, if the factor of authenticity is so crucial for observation and description as I am suggesting, then linguistics and applied linguistics should brace themselves for a major scientific revolution whose repercussions will inevitably be felt in TEFL. In exchange, TEFL can offer our best resources for measuring the applicatory adequacy of new theories.

The upcoming 'paradigm shift' can be predicted to transform the entire concept of a language: not a *static system of units* (phonemes, morphemes, phrases, sentences, etc.) but a *dynamic system of relations*. Instead of a dichotomy between 'langue and parole' or 'competence and performance', we can recognise a *dialectical cycle* between *combinability* (language as potential system) and *combination* (text as actual system). And instead of separating a 'grammar' of 'rules' from a 'vocabulary' (or 'lexicon') of 'words', we can explore the *unified lexicogrammar* for the typical grammatical combinations called *colligations* and the lexical combinations called *collocations*.

Some of these conceptions and terms have been with us for quite some time, but their 'revolutionary' impact centres on fully recognising authenticity to be the fundamental precondition and constant requirement for observation and description to achieve applicatory adequacy. Authenticity is first of all an empirical property of data certified by their occurrence in a context of situation. But authenticity can become a property of an applicable theory or description only if some challenging problems can be solved.

I should emphasise at once that these problems do not arise from principled weaknesses inherent in corpora, despite what is consistently alleged by those who oppose the use of corpus research in applied linguistics in language teaching (e.g. Widdowson 1991). Rather, they are problems which have been inherent in language research and language teaching all along but which corpus studies allow us to recognise and formulate. The corpus raises questions rarely posed, let alone answered, in mainstream linguistics, such as:

(a) How big is a language?
(b) What is its ratio between uniformity and diversity?

(b) What is its ratio between regularity and accident?
(d) How much data is enough for the description of a language?

Instead, linguistics has propagated reassuring abstractions and generalities, e.g.: 'language' is 'like a dictionary of which identical copies have been distributed to each individual' (Saussure, 1966 [1916]: 19); or 'linguistic theory' is 'concerned' with a 'completely homogeneous speech-community' (Chomsky, 1965: 4). Language is declared to be uniform without even bringing forth authentic data. Corpus studies deny us this easy reassurance.

I would make a similar point for applied linguistics and TEFL. The problems in setting up and applying corpora reflect prior indecisions about the world-wide mission and audience of EFL insofar as these have been sustained by using non-authentic English to project a simplified uniform vision of the language to be taught the same way to everybody, everywhere — and from the same textbooks. Now that corpora are accessible to the teaching and learning of English for specifically 'academic and professional purposes', we find ourselves perplexed by the shift to authentic English. The shift is so difficult not because (as some have argued) corpus data do not represent the English language, but because the concept of 'authentic data representing a language' has not yet been operationally defined in language pedagogy.

Yet we may achieve major progress through a principle we might call *dialectical resolution*, whereby the problem arises from the same source that will lead toward its eventual solution. Corpus studies have been exposing problems that remained implicit or excluded in non-corpus studies of language; but the corpora themselves provide the raw materials for finally resolving those problems.

The toughest problem is unquestionably the *diversity* that comes with authenticity. How far is any one set of data relevant to another set in the corpus, or to the corpus as a whole? Which sources of data deserve to be represented, and in what proportions? In theory, diversity might be mastered by breadth. At each progressive jump in size, the distinctions between accidents and regularities will become more precise; some presumed accidents will turn out to be regularities, and vice-versa. Sinclair (1999a) has recently aired the prospect that the 'generic or reference corpus', such as the Bank of English, 'currently approaching 500 million words', will be 'large enough to smooth out many of the idiosyncrasies

of individual authors and texts'. But of course the concept of 'idiosyncrasy' is itself an unsolved problem within the larger problem of diversity.

A brief demonstration may be helpful here. In a previous analysis of the uses of the Verb 'warrant' (Beaugrande 1996), I analysed the 228 occurrences in the Bank of English (then at 220 million words) and found just 4 in colligation with First Person Subject, used when you want to indicate you feel sure about something though you can't point to actual facts. Now, my 10-million-word corpus of British and American writers returns 198 occurrences, of which fully 75 show this colligation. The proportions seem dramatic until we notice that 25 of those are from Fielding's *Tom Jones* alone, e.g.:

[28] I warrant you will never repent having the money into his hands.
[29] Why, you thought, sir, I knew nothing of the matter, I warrant you, about Madam Sophia.

This corpus displayed an authorial idiosyncrasy which would be mistaken for a regularity if we did not attend to our sources.

Again in theory, breadth of observation should move on a stable upward curve as data are added; but the issue in fact hinges on how new data relate or compare to previous data. Since the 'generic corpus' should reasonably seek out as much diversity as is representative of authentic English, its significance or information value gains little from adding more data of same type. This problem applies especially to mass media, such as the plentiful newspapers conveniently posted on the Internet. Their diversity as data is restricted in being authored by a relatively small, well-trained group of writers, and being edited by an even smaller group. The total effects of these restrictions upon the representative qualities of the data are yet to be assessed. I would also note the massive frequencies I found in the BoE in July 1994 of key-words like 'death', 'kill', 'murder', 'massacre', 'shooting', 'robbery', and 'rape', reflecting the morbid interests of mass media more than the frequencies of authentic English at large.

Evidently, breadth of coverage does not match the size of the corpus, but must be factored out between size and diversity in relation to authentic English. Here, dialectical resolution could apply: having admitted the problem, we use corpora to get it under control. Daunting labours await us in accounting for our intuitive distinctions among the English used by prominent speakers or writers (e.g., politicians, novelists,

columnists), or by media (e.g., newspapers, magazines, chat sites), or by professions (e.g. doctors, lawyers, scientists). No doubt we will discover factors that render the use of corpora in language teaching more complicated but also better secured.

This point applies most trenchantly to the use of literature. Literary texts are widely used in EFL programmes without an operational account of the relation of literary English to other Englishes, particularly to those of the learners. Since current definitions of style in stylistics accentuate the personal and special qualities, literary English should surpass all others in its diversity, and also in its remoteness from non-authentic English. Quite plausibly, learners face a daunting jump in order to access literary English without adequate fluency in authentic English.

Again, dialectical resolution could apply. Learners could work with a series of corpora of authentic English expressly designed to promote the fluency needed for literary English, which could in turn be approached through a series of literary corpora arranged in terms of their accessibility. Some principled decisions would be required about which literary texts or text types merit study within the limits of a given programme format. Yet that this approach could materially increase the range of literature we could effectively cover in practice.

If the breadth of an applicable description creates tough problems, depth creates even more. Already at this stage, we can see the hopelessness of any cut-off in depth analogous to the cut-off in breadth derived from relative frequencies in such applications as dictionaries. How deep a description should extend will fluctuate sharply across a very large corpus. Frequency of occurrence is by no means a reliable indicator of appropriate depth, though some working ratio may eventually be determined.

For depth, our greatest problem is undoubtedly our *categories*. The available categories in linguistics and EFL are a mix of traditional Latin-based grammar (e.g., 'indicative', 'intransitive') with various modern approaches, sometime descriptive (e.g., 'constituent', 'verb phrase') sometimes generative (e.g., 'second language acquisition', 'universal grammar'). Not surprisingly, they do not constitute a unified description either in theory or in practice; and corpus studies keep turning up gaps.

I shall end this section by illustrating one noteworthy gap. So many combinations in authentic English carry attitudes of good and bad, or pleasant and unpleasant, or approval and disapproval, as to constitute what I proposed to called 'standards'. No terms are widely established;

we can't use 'positive' (which could mean certain) and 'negative' (which is a category of the Verb Phrase). I have chosen 'ameliorative' and 'pejorative' which, though also from Latin and a bit academic, can be reserved for attitudes alone.

One class of relevant data are Adjectives that do not supply the attitudinal quality of their Head Noun, but only highlight a quality the Noun would imply by itself (cf. (Tognini Bonelli, 1993: Sinclair 1999). In my corpus of British and American writers, 'beautiful' often occurs with things I at least would not expect [30-34], whereas 'ugly' often occurs with things that could hardly be otherwise [35-39] and never with anything I'd expect to be beautiful.

[30] One night there was a beautiful electric storm (Willa Cather)
[31] My heart quite fails me when I think how I might have lost that beautiful luncheon-basket. (Kenneth Grahame)
[32] 'Hound never ran on a more beautiful scent', responded the scout, dashing forward (Fenimore Cooper)
[33] We must go and visit our beautiful suburbs of London (Thackeray)
[34] I come quite over-powered. Such a beautiful hind-quarter of pork (Austen)
[35] he opened his jaws, rolled back his lip in an ugly snarl (Zane Grey)
[36] what an ugly monster it was! Only his horned head belonged to a bull (Hawthorne)
[37] with a bang an ugly black imp appeared and croaked a reply (Alcott)
[38] 'He were an ugly devil', cried a third pirate, with a shudder; 'that blue in the face, too!' (Stevenson)
[39] it was as ugly gaping wound as surgeon ever saw; more than two feet (Melville)

'Beautiful' was five times as frequent, and many occurrences were collocated with things that could be or not be so, the most numerous, perhaps inevitably, being 'girl - lady - woman' — in the discourse of fiction, after all.

In corpus data, the same word may be found to carry both attitudes. In some collocations, the Adjective 'serious' has the ameliorative meaning 'significant' or 'sincere' [40-43], and in others the pejorative meaning

'grave' or 'alarming' [44-46]. If the Noun is vague, like 'thing', the pejorative dominates [47-48].

[40] She did it constantly, with such a serious enthusiasm that he grew fond of watching her
[41] The discussion carried me far afield in perhaps the most serious economic reading I have ever done (Jane Addams)
[42] If Mr. Goodwood were interested in Isabel in the serious manner described by Miss Stackpole he would not care to present himself at Gardencourt (Henry James)
[43] you have actually given her reason to flatter herself that you had the most serious designs in her favour. (Fielding)
[44] The English Government took an extremely serious view of the matter (Strachey)
[45] something has occurred of a most unexpected and serious nature; but I am afraid of alarming you (Austen)
[46] I found the undertaking even a more serious task than my fears had led me to imagine (Poe)
[47] the first time life ever struck Jones as a really serious thing was when the Dean told him he must leave school. (W.E.B. DuBois)
[48] To some it seemed that now that they were in actual possession of it, freedom was a more serious thing than they had expected (Booker T. Washington)

As remarked in the source text for [48], 'freedom' can seem pejorative to slaves who are too old or weak to find new jobs.

In this section, I have tried to show that advocating the introduction of corpora as authentic English into EFL programmes in no way implies a brisk optimism that overlooks substantive problems in the relation between theory and practice. By themselves, the practical benefits should be easy to grasp. TEFL shifts from a frontal, teacher-centred exercise in the production of correct though non-authentic sentences over to a learner-centred joint exploration and appreciation of authentic texts. Teachers are freed from tedium of inventing and writing out data, and from any residual anxieties about their own command of standard English. And even everyday classwork can lead to the discovery of subtle and previously unnoticed 'standards' as a tangible and creative achievement.

But in the wider context, these benefits need to be secured through strenuous theoretical work. In this exploratory stage, those of us who

are most involved with corpora in teaching are also the most compelled to assess the problems. In contrast, people who are not involved and yet publicly oppose the use of corpora on some abstract principles are the least qualified to do so. The burden of proof should rest upon them to demonstrate how exposure to non-authentic English can nurture fluency in authentic English.

Pratical corpus work for students

The United Arab Emirates as a whole is a patchwork of ESL and EFL environments. The region was a 'British Protectorate' from 1820 to 1971, but, prior to the discovery of oil in 1958, the British presence was minimal. No attempts had been made to establish British schools or cultural centres, nor to encourage the spread of the English language. Today, the larger towns and cities are strongly ESL, but mainly among the enormous communities of expatriates from Asia, Europe, and America. Among the actual UAE citizens, who constitute between a fourth and a third of the total resident population (1996 estimate), Arabic is clearly the dominant language.

The student body of the United Arab Emirates University is almost entirely composed of UAE citizens, aside from a small contingent from other Arab regions, such as Egypt, Jordan, and Syria, whose families are employed here. Apart from the modest portion who have attended English-medium schools, the fluency level is closer to the EFL than the ESL environment.

To increase their exposure to authentic English, our students are using WordPilot©, a corpus-based resource program developed by John Milton at the Hong Kong University of Science and Technology (Milton, 1999). It can be used to access on-line displays of authentic usages for specific expressions and combinations while reading or writing in a programme like WORD. Having selected a suitable corpus, such as 'British academics', students can click on a doubtful word in their own text and see some authentic examples. They can see, for instance, that people 'differentiate' not by separating mixed substances like 'milk from water' in sample [9], but by developing or recognising a quality that identifies one kind as distinct from another, as in [49-50].

[49] This interest in work differentiates the workman from the criminal (Veblen)

[50] Shakespeare differentiates his heroes from his villains much more by what they do than by what they are (Bernard Shaw)

A regular exercise in my 'semantics' course is to query a key-word and explain the different meanings it can assume in context. For example, here are some student responses to the key-word 'fine':

[51] he tried to raise her self-respect with fine clothes and flattery (Emily Brontë) => elegant
[52] Harriet was short, plump, and fair, with a fine face, blue eyes, light hair (Austen) => delicate
[53] I heard rain strike earth in fine needles of water (Joyce) => thin
[54] Mr Elton has not such a fine air and way of walking as Mr Knightley. (Austen) => dignified
[55] 'A fine husband you are!' said Mrs Glegg scornfully. (George Eliot =>) worthless
[56] Things are come to a fine pass when one sister insults the other! (George Eliot) => dreadful
[57] Don't trust them fine-talking men from the big city. (George Eliot) => smooth, flattering
[58] I shall come and see your mother some fine day. (Alcott) => some indefinite future day
[59] we get a fine day, and then down comes a snapper at night. (Thomas Hardy) => sunny

Most students were not familiar with the collocation 'fine day', and, in an amusing cultural contrast to British usage, some guessed it must be a cool and cloudy day — an apt guess here in one of the hottest desert regions on earth, where sunny days reach 50° Celsius. They also did not know what a 'snapper' could be in [59], nor did I — surely not a nocturnal fish of the species *Lutjanidae*. We did not find it again in the corpus, but we did find 'cold snap', e.g.:

[60] he closed my carriage door one sleety day during the cold snap of February ninety-three (Joyce)

A 'snapper' would be a regional variation among Hardy's rustic farm hands.

The collocation 'fine day' was found to have subtle social functions when we queried it in the corpus, rather like zooming in on some detail of a visual scene with a camera lens. These data proved helpful:

[61] One day Parson Thirdly met him and said, 'Good morning, Mister Everdene; 'tis a fine day!' 'Amen', said Everdene, quite absent-like, thinking only of religion when he seed a parson. (Hardy)

[62] there stood the Queen in front of them, with her arms folded, frowning like a thunder-storm. 'A fine day, your Majesty!' the Duchess began in a low, weak voice. (Carroll)

[63] 'Good morning, Mr. Watty; it's a fine day for walking, isn't it?' Seeing that the stranger still lingered, Mr Lowten shut the door in his face. 'There never was such a pestering bankrupt since the world began, I do believe!' (Dickens)

Quite irrespective of the weather, you can greet somebody with 'it's a fine day' in order to appear sociable and agreeable, though you may not get the effect you wanted [61]. As useful variations, you can use the greeting to mollify a hostile encounter [62] or hint that an unwelcome encounter should come to a speedy end [63].

Another interesting exercise is to examine the meaning of major terms within a single text. In Jane Austen's *Pride and Prejudice*, we searched and sorted out all collocations of 'pride' and 'proud'. These ones concerning Mr Darcy began to suggest a pattern:

[64] everybody says that he is ate up with *pride*

[65] He is not at all liked in Hertfordshire. Everybody is disgusted with his *pride*.

[66] he was discovered to be *proud*, to be above his company, and above being pleased;

[67] His character was decided. He was the *proudest*, most disagreeable man in the world.

[68] It is wonderful, — replied Wickham, — for almost all his actions may be traced to pride.

[69] she tried to remember something of that gentleman's reputed disposition, when quite a lad, [...] and was confident at last that she recollected having heard Mr. Fitzwilliam Darcy formerly spoken of as a very *proud*, ill-natured boy.

We noticed how Darcy's pride was typically stated as circumstantial opinions of unidentified persons, such as 'everybody' [64-65]. Just because he failed to be sociable at one assembly, his proudness was instantly and irrevocably 'discovered' and 'decided', where the colligation with Passive Verbs and the collocations 'ate up' and 'most...in the world' subtly deconstruct the force of what is asserted as absolute

truth [64-67]. Then we encountered the personal testimony of Mr Wickham, whose sincerity we knew to be worthless [68]. When Elizabeth repeats Mr Wickham's account, the circumstantial quality becomes too elaborate to be easily ignored: Mrs Gardiner rummages in her memory and 'at last' dredges up a foggy 'recollection' of another unidentified opinion about a 'reputed disposition' [69]. The data led us to see how the reader is positioned to believe Mr Darcy proud by getting implicated in the 'prejudice' harboured against him by Elizabeth Bennet, whose perspective is subtly entwined with that of the ironic narrator.

No doubt the students could have dug this insight out the professional literary criticism on the novels of Jane Austen. But they learn and enjoy much more by working it out themselves from the actual language of the text. They have not merely produced a key to appreciating this particular novel and its epigrammatic title. They have hit upon a practical object lesson in the technique of irony. And, they have discovered ways to appropriate literary English despite its remoteness in time and place.

Pratical corpus work for teachers

For the present, EFL can provide indispensable practical assistance in identifying the modes of data in and about authentic English that corpus studies could reasonably provide. Indeed, progress toward applicatory adequacy can be achieved only through a sustained co-operation of corpus studies with language pedagogy.

One familiar problem arising in a course in 'English grammar', put to me by colleagues at an African university in December 1998, prompted me to consult the COBUILD (Collins Birmingham University International Language Database). The database, popularly called the 'Bank of English', is the world's largest computerised data corpus, then containing 329 million words of running text, of which 20 million were spoken data. It draws upon a range of contemporary spoken and written sources, including: British and American books; newspapers *(Times, Independent, Guardian, Today, Wall Street Journal, New Scientist, Economist);* magazines (e.g., *Esquire, Good Housekeeping);* ephemera such as letter-box mailings (e.g., YMCA appeal for homeless people, Friends of the Earth Tropical Rainforest Campaign), radio broadcasts (British Broadcasting Corporation in the UK and National Public Radio in the US); and recordings of conversations.

The problem in English grammar concerned the variation in usage between '*if I was*...' and '*if I were*...' followed by a Noun or Pronoun. Some colleagues insisted that: 'if I was' is incorrect, and only 'if I were' is correct. The advice I found in the voluminous *Grammar of Contemporary English* and the *Comprehensive Grammar of the English Language* proved ambiguous. This 'were' is described there as the 'singular past subjunctive form', sustained 'by convention' for 'the idiom "if I were you"' (Quirk, Greenbaum, Leech, and Svartvik, 1972: 747; 1985: 1094). However, those same *Grammars* concede: 'the subjunctive is not an important category in contemporary English and is normally replaced by other constructions' (1972: 75); and 'the subjunctive in modern English is generally an optional and stylistically somewhat marked variant of other constructions' (1985: 155). If so, then what remains of the subjunctive is preserved by certain colligations and collocations — by standards rather than rules.

The 1972 *Grammar* proposed a distinction between 'hypothetical conditions conveying the expectation that the conditions will *not* be fulfilled' versus 'open conditions leaving unresolved the question of the fulfilment of the condition' (1972: 747, their emphasis). The 'open' ones 'have also been termed "real" or "factual"', and the 'hypothetical' ones '"unreal" or "counterfactual"' (1985: 1092). The '*were*-subjunctive' was said to be 'restricted to one form' and to be 'hypothetical in meaning' (1972: 77). Yet ambiguity emerged again in the warning that 'both the past subjunctive and the past indicative are possible for hypothetical conditions' and are 'occasionally used in formal contexts', as in '"If it was/were to rain, the ropes would snap"' (but is that formal?); still, 'the subjunctive is preferred by many in formal contexts, especially in formal written English' (1972: 748; 1985: 1093f). We shall see the corpus data telling a different and more interesting story.

The Bank of English returned 2061 lines for 'if I were' and 2876 for 'if I was'; at least both usages are truly alive and well. For purely practical reasons, I decided to start with examining and classifying roughly 10% of these[7]. During this work, I noticed that '*if I were you*' appeared in 20 lines, and '*if I was you*' in only three lines, whilst other usages with Pronouns after 'were' or 'was' were quite rare. To check these proportions, I requested all lines from the Bank of English where 'if I were' and 'if I was' were followed by any of the Pronouns 'you, he/him, she/her, they/them'. This time I got back 402 lines, and sure enough the

frequencies across the entire corpus were drastically uneven. No less than 282 lines attested 'if I were you', whereas 'if I was you' trailed at 37. The rest, those having Third Person Pronouns, were at best marginal, some hovering between 10 and 20 and some close to or equal to zero. After eliminating a few false alarms[7] (e.g. 'if I was her dog'), I got these totals:

if I was he	0	if I was she	2	if I was they	0
if I was him	17	if I was her	5	if I was them	11
if I were he	3	if I were she	0	if I were they	1
if I were him	18	if I were her	6	if I were them	10

These figures indicate that a usage commonly recommended for 'standard English' in EFL textbooks — 'if I were' + Subject Pronoun — is no longer secured in authentic English. The old Subjunctive 'were' is surviving much better than the presumably standard Subject Pronoun after it, and was found to colligate with the Object Pronoun roughly as often as did the Indicative 'was'.

The data also indicate that the applicable standard is not the distinction between 'hypothetical' versus 'open conditions' proposed in the two *Grammars* cited above. To be precise, *all* cases are 'hypothetical', since 'I' can never be anybody but 'I'. Nor did the data show 'were' being preferred over 'was' when the prospect that 'I' might be another person was particularly improbable. For example, I found both forms for scenarios of grandly imagining to be one of the Royal Family [70-71], but also for ones of prosaically imagining to be one's own sister [72] or a worker for another company [73].

[70] In regard to Diana, Joan Collins offers this suggestion: 'If I were her I would come out here to LA, hire the biggest agent and get $25 million to do one film.'

[71] he wanted to tell the prince 'what a fool he was to let Diana go'. He said: 'She is a beautiful woman and my favourite. If I was him I'd beg her to come back.'

[72] I was surprised because my sister is not the submissive type of wife who obeys whatever her husband says. I thought if I were her, I would just put my kids in the car and not care what my husband says.

---

[7] These false alarms explain why the figures given here do not add up to the full 402.

[73] Tim, a construction worker who supports Target, shook his head in disgust. 'If I was them, I would go somewhere else'.

The data also indicate that being 'preferred in formal contexts' cannot be the applicable standard. I found it in numerous contexts which were decidedly informal, e.g.:

[74] If I were you, I'd move on this real soon and come up with something, or you're going to be too late.

[75] 'A young chickabiddy like you's not done for yet. You know what I'd do if I were you? I'd make a pact with myself to succeed to spite the beggars!'

[76] He drank eleven shots before he could feel the influence of the alcohol. He ordered his twelfth and the bartender told him: 'If I were you, I'd get some air'. 'I can pay you', Cross told him. 'That ain't the point', the bartender said.

The applicable standards are rather to be found in the social functions of real-life discourse. Among the major functions we find a triad of closely related *discourse moves* or (to use the older term) *speech acts*. *Advice* is given when the speaker (or writer) has the friendly intention of suggesting what the hearer should do. A *warning* is issued when the speaker has the (more or less) friendly intention of pointing out potential bad consequences for the hearer. A *threat* is made when the speaker has the unfriendly intention of frightening the hearer and coercing some action to be done or avoided.

In some data, we can clearly distinguish which of these three discourse moves or speech acts was intended, e.g., the advice in [77-78], the warnings in [79-80], and the threats in [81-82]. Notice again the lack of 'formality'.

[77] 'I'd get some sleep if I were you. You'll need to be up at six to catch the early morning flight from Heathrow.'

[78] If I were you I'd keep pestering them. Because sooner or later a job will come up.

[79] The builder looked at it and said, 'I hope you're not thinking of filling that thing with water. I wouldn't if I were you — it'll go through the floor.'

[80] 'Colonel Sharpe won't dare kill you in my ballroom, because I won't let him. But if I were you I'd give him his wife back and find yourself someone more suitable.'

[81] I wouldn't come home if I were you. You should stay away from my patch. There are people who know that you grassed[8] and if they know where to find you

[82] 'The blood of the mob is up! If I were you, I'd clear out of town now with as much as you can carry, because you've been found out!'

Sometimes the move was explicitly named:

[83] 'Well, if I were you', she adds by way of some unsolicited advice, 'I'd watch out for that girlie of yours.'

[84] 'I wouldn't go in there if I was you', warned the young man in the office. 'You've no idea what fish meal smells like when it's being dried.'

[85] 'You're a fortunate man, but becoming a real irritant. I wouldn't put too much faith in that chain, if I were you.' 'If you threaten me or use any force, I shall inform the police.' 'They might be a while getting here'.

For the hearer, the advice may not be much appreciated [86]; or the warning or threat may not have the intended effects [87].

[86] when your having your therapeutic whinge about your last date from hell they don't just listen sympathetically, they wade in with remarks like, 'If I were you...' and 'Once you've really experienced love, like me, you'll realise...' Why won't they just shut up?

[87] 'Gordon's going to want your ass in a sling for this one. If I was you, Wade, I'd move to Florida. Tonight.' 'But you're not.' 'Nope, I'm not. Thank Christ.'

Perhaps all three discourse moves risk being perceived as irritating presumptions that I know what's good or bad for you better than you do. If so, saying 'if I were you' or 'if I was you' lessens the risk by seeming to interchange roles and implying: 'I'm not telling you what to do, I'm just telling you what I would do if I happened to be in your place'. The two colligations could thus offer some means of saving face for the hearer who heeds the move with actually being commanded, or for the speaker who gets ignored or challenged. Further face-saving might be achieved with such reservations as 'you know what I'd do?' [75], or 'I hope you're not thinking of doing that' [79].

As we might predict, the contexts and situations differed when the item following 'if I were/was' was *not* 'you'. The speaker can still

[8] grassed: gave information to the police (*COBUILD* 634)

issue advice, warnings, and threats, but without having to address the intended recipient. We might call these moves *playing out a scenario*: freely imagining what would be the case if the speaker were in somebody else's place, however fanciful — say, if you were a consultant 'advising the Government' of Britain [88], or the American president forming the 'cabinet' [89], or the 'prime minister' of Australia 'negotiating the budget' [90].

[88] Once people have realised the tunnel is still there and there's the chance of a price-cutting bonanza, they will see that its ownership doesn't matter. If I was advising the Government I would tell them to tough it out

[89] Interestingly enough, if I was choosing him for a Cabinet slot, I would have put him at HUD, Housing and Urban Development. Instead, Clinton's putting him in Agriculture.

[90] I can only imagine the outcry if I was prime minister and I was negotiating the Budget with the Business Council of Australia a week before the Budget was due to be brought down.

Conspicuously popular was advice given about sports. You can play out a scenario of being the star racing 'driver' Michael Schumacher deciding where to 'sign' [91], or the star boxer Mike Tyson doing something as far out of character (for him) as 'reflecting upon life' [92]. Your scenario can even elevate you into being the 'football coach' not just for Brazil but for every team in the whole world [93].

[91] 'Ferrari's decision to sign Schumacher is the right one', Prost added. 'He's the best driver; however, if I were he, I would have stayed with Benetton for another year.'

[92] But the Briton believes Tyson should not be in too much of a hurry to lace up the old gloves again, and warned: 'If I was him and had spent so long inside I wouldn't rush into anything. I'd want to reflect on life and enjoy my freedom.'

[93] I was criticised when I said if I was coach of Brazil I'd make them a better team but I only said that because I believe my way is best. My style is not only suitable for Norway but football everywhere.

For the English-speaking press, these wishful scenarios serve the function of keeping the immense audience of sports fans listening or reading during the times when actual sports events are not in progress.

The audience can play out their own scenarios and identify with their heroes by hearing how other people do so.

A more prosaic major source of scenarios in the COBUILD data was family matters with their many expectations and deliberations about what one could or should do, e.g., when your 'sister is not the submissive type of wife' back in sample [72]. Families seem preoccupied with making sure that 'family life goes on, no matter what' [94], e.g., when people are 'feeling awful' or 'ashamed' or having 'rows every day' [95-96], or when parents are trying to control their adult children [97].

[94] I couldn't cope with it if I was the hysterical type. But like most women, I have made my mind up that my family life must go on, no matter what

[95] These aren't very kind thoughts of mine, particularly as I know how awful I'd be feeling if I was her. Her life revolved round my father and she now wants it to revolve round me.

[96] my sister she's never once had a holiday because he didn't earn a decent wage. I told him I'd be ashamed of that if I were him. After that there were rows every day and when we weren't rowing we didn't speak at all.

[97] [female speaker:] twenty-three and his parents won't let him come and he [male speaker:] God / that's awful [female speaker:] abides by that / if I were him I'd just say I'm going out [male speaker:] hitch into Stafford and get on a train / [female speaker:] yeah

As conceded by the *Grammars*, the distinction between 'hypothetical conditions' versus 'open conditions' certainly does not match the distinction between subjunctive and indicative. The Bank of English data displayed a subtle range of ways for indicating that something is or is not the case, or might or might not be, and so on. Usually, these matters were decided by the context but sometimes were strategically left undecided.

'Open conditions' were well attested with the implication 'if it was the case, as it well might be':

[98] 'Do you think that having a personal mobile phone would increase your feeling of security?' 'Yes, <u>if I was out alone at night</u>'

[99] I told her that I should be back by the end of the week with any luck, and that I would communicate any change of plan to the office <u>if I was delayed</u>.

In other contexts, the implication was: 'if it was the case, and in fact sometimes it was':

[100] if I was drinking with mates, I had to drink them under the table as a matter of principle. I would do anything or try anything to show how big I was.

[101] he was physically abused when he was young, but was not aware it was illegal: 'I thought if I was cruelly treated, if I was tortured, maybe it was right, maybe it happened everywhere'.

In a few contexts, the implication was evasive: 'if it was the case, and maybe it was':

[102] Geography it's okay like but I could go home [...] and get out my sister's worksheet you know like if I was having problems

[103] I knew that Mike was highly sexed and if I was not giving him enough he was going elsewhere for it. Deep down I felt guilty that I didn't want him more sexually; that the fault was mine, not his.

In still others, the implication was: 'if it was the case, but there was good reason to doubt it':

[104] If I was so all-fired bright, as my parents, who had patently no basis for comparison, seemed to think, why did I have to keep learning this same thing over and over?

[105] it also showed the persistent attitude on the part of the Home Office towards me. If I were such a high security risk, why had I not done anything all that time when I was out on bail awaiting trial? If I were really in the IRA, why had I turned up of my own accord?

Quite frequently, the context indicated: 'if it were the case, but it certainly isn't or wasn't':

[106] When I made out the cheque for £60 they went with the number rather than my name. If I were a dishonest person I could have gone through the cheque-book making out cheques and they'd all have been debited to this other person's account.

[107] If I, the chief city official, was not seen on the dais viewing the parade, it would be a catastrophe. This simply couldn't happen.

[108] She is thinking about pursuing the matter through the courts, but is disillusioned by the response: 'If I was a member of the Royal Family people would care. But I am Mrs Nobody, Mrs Ordinary from Milton Keynes.'

In infrequent contexts, the implication was like a commitment: 'if it was to be the case, something had to be done, and so it was', and I there found only the colligation 'was' plus 'to' or 'going to':

[109] I didn't know what would happen if I let him go but I knew I had to let him go if I was to restore harmony with the state and with my employees.

[110] if I was going to break that destructive cycle I'd inherited from my mother, it needed something as extreme as that to do it. It was very painful, but it was right.

In plentiful data, '*if*' was equivalent to '*whether*' when collocated with a preceding expression of uncertainty like 'ask' [111], 'wonder' [112], or 'not know' [113]. The context often indicated why somebody might be uncertain, as in [111-112], and whether it was the case as in [113], or was not the case, as in [114]. But in a few data, the uncertainty remained strategically unresolved. Could 'depression borrowed' from Philip Larkin plunge you into a 'nervous breakdown' [115]? Was our hero 'being watched' by gunmen on the 'roof' after all [116]?

[111] One day Mrs Luppin remarked that I was looking a little off-colour and asked if I was feeling all right. I told her about my sickness.

[112] then I got pregnant and then as time got further along, I started getting a little scared, wondering if I was going to be able to do things right

[113] I didn't know if I was going to do this assignment [...] But late last night, I decided to.

[114] I've never seen chauvinism like it is here. The judge was out of order. [...] He asked if I was the kind of girl who would take off her top and bra in a roomful of people after a few drinks.

[116] I wrote seven poems to enter, very much borrowed from the Philip Larkin style. I borrowed his sense of depression, too, so all my friends were phoning to see if I was having a nervous breakdown.

[117] No one with a gun had shown himself above the roofline but how could I tell if I was being watched?

Among the 19 occurrences of 'ask' and the 7 of 'wonder' in this usage, only one each chose the Verb 'were', and in both cases with the implication that it was *definitely not* the case:

[118] At each store, when I asked if their beds came assembled, an

assistant asked if I were disabled or infirm. Only then might the delivery men put them together

[119] I stayed away from girls for a long time. I wondered if I were homosexual. But I knew I wasn't

Projecting this rare colligation into the Present Tense could yield the 'classic instance of hypercorrection' cited in the *Comprehensive Grammar* as 'the pseudo-subjunctive in "I wonder if he were here"' (1985: 14)[9].

The Verbs 'say' and 'tell' occasionally appeared in cautious collocations which stopped short of actually 'saying' and 'telling' and merely invoked a scenario of doing so:

[120] If I were to say that the English ICA has taken a sensible step in setting up a working party to look at the issues, members could be excused a hollow laugh.

[121] And if I was to tell you that Lydia was the most extreme, the most uplifting yet simultaneously depressing, artist and performer and actress I've ever seen, how would you react?

This move can also save face when you are concerned about how people might react to what you would say.

The data for usages with *'as if'* indicated still other functions. The routine function was 'as if it were true, but of course it wasn't', as became obvious by other choices of words in context:

[122] 'Françoise, we flew to the Moon!' She looked at me as if I were completely insane. 'What did you say?'

[123] When I tried to meet them another morning they had evidently been told that they must not speak to me, and only little white faces looked at me as if I were an evil ghost.

But I also found data shading over into the implication of 'it did seem as if it were true, even though it wasn't':

[124] I travelled by train from Montreal to Toronto. All my baggage was booked in at the station and taken by trolley to the luggage car, just as if I was flying.

---

[9] But compare Jane Austen: 'she asked the chambermaid whether Pemberley were not a very fine place' (*Pride and Prejudice*).

[125] I was beginning to feel as if I were being subjected to repeated demonic attacks, but of course this was nonsense.

The next shading had the implication of 'it did seem as if it were true, and maybe it was':

[126] I always felt that people looked at me as if I was a little bit rebellious and a little not the conservative person.

[127] I felt as if I were just going to school to fill / you know / fill some time in / I didn't feel as if I were going for a purpose / you know / it were just somewhere to go

The final shading had the implication of 'it did seem as if it were true, and actually it was':

[128] It gave me a strange, shy feeling, as if I were in an intimate situation with strangers. But then, that was where, I thought, I was.

[129] My gentleman repeatedly looked down at me, and, so I thought, gesticulated in my direction as if I were the subject of a heated discussion. At last Mr. Grummage came down to the dock. [...] his face was flushed, with an angry eye

This overview of data was relayed to my TEFL colleagues who were arguing about the correctness of 'if I were...' versus 'if I was...'. The data gave them a much sharper picture of the actual standards based on the relevant social functions in typical contexts.

Such an experience may be a general outcome of examining authentic data. Teachers can raise their sensitivity for the multiple standards that apply to actual usages of present-day English. Here are some plausible candidates:

(1) the standard regularities in the grammar, such as the tiny choice of remaining forms for distinguishing between indicative and subjunctive;
(2) the resilience of some parts of a standard grammatical colligation while the rest is worn away, e.g. 'if I were you' remaining authentic long after 'if I were she' or 'if were I they';
(3) the social functions of some collocations, e.g. 'if I were you' for saving face when advice or warnings are being given;
(4) the adaptations within contexts, e.g., indicating the probability of the scenario introduced by 'if I was/were' being the case or not, as

when my 'drinking with mates' was certain [100], whereas my 'flying' was impossible when 'travelling by train' [124];

(5) the meanings of specific expressions affecting this probability, e.g., my being quite capable of 'doing this assignment' [113] but not of being 'completely insane' [122].

(6) the social standing or the speaker or writer, e.g., whether the scenario of being one of the Royal Family is played out by Hollywood soap-opera star Joan Collins [70] or 'Mrs Nobody from Milton Keynes' [108].

(7) the topic of the discourse, e.g., whether you're talking about an exciting sport like car-racing and creating scenario of being a famous hero like the dashing 'Schumacher' [91], or just about humdrum family life and imagining how 'awful' my relatives must be 'feeling' [95].

(8) the text type, e.g., novels of crime or espionage where people often threaten each other to 'clear out of town' [82], or promotionals dressed up as opinion surveys about the benefits of 'having a personal mobile phone' [98].

(9) the modality of speech or writing, which may lead toward spontaneous usage for speech, e.g. [97], and more careful usage for writing, e.g. [129]; our data were often in between due to the frequent use of written discourse to represent speech, e.g. [74-75] and [81-82].

Such lists can only be provisional this stage, and certainly cannot qualify as scientific discoveries. The diversity of the data sample undoubtedly falls far short of the diversity of contemporary usage, despite the huge size of the total corpus. But the diversity is undeniably greater and richer than could be represented by any sampling of non-authentic data, however large. In exchange, this much diversity already entails some tough theoretical problems, as explained in section 2. As a practical heuristic in place of the unwieldy documentation of hundreds or thousands of samples, we can use the data themselves to infer the various styles, registers, or genres.

Nor again can such a presentation of data for EFL teachers purport to be a 'linguistic description' satisfying the criteria of science for 'descriptive adequacy'. But it can be fairly assessed for its applicatory adequacy if the teachers find it useful for their own practices, particularly for 'communicative' method, where authentic data should be heartily welcome.

Back to the future

This paper may have provided some grounds for guarded optimism regarding the uses of corpora of authentic English in the teaching and learning of EFL. At this stage, we should be wary of untested claims either for or against. Authenticity will not come cheaply or reassuringly to linguistics and applied linguistics after a long period of reliance on non-authentic data. But perhaps this very reliance has been unaffordable in quietly undermining or defeating our theoretical and practical projects, like an invisible barrier interposing non-authentic language and blocking our vision of authentic language.

In the new millennium, the most massive and vital project turns out to be providing access to English as a Foreign Language. Teachers and learners of EFL simply cannot afford to placidly wait for some future time when every problem in theory and practice has been solved. Without their active participation, without their input and feedback at every stage, those problems cannot be solved at all.

Perhaps my use of the term 'scientific revolution' in section 2 sounds unduly dramatic. In hindsight, the familiar illustrations in the philosophy of science, such as the discovery of oxygen, X-rays, and the Leyden jar in Kuhn's (1970) exposition, seem inevitable, driven by theoretical anomalies and practical applications that soon became indispensable to modern technologies. Language studies have been perhaps too adept in camouflaging or dismissing our theoretical anomalies and diluting or postponing our practical applications, until we find ourselves overtaken by technologies coming from outside. Now the challenge assumes an unprecedented urgency:

> Language teaching is an application [that] has not as yet attempted to derive its authority from technology, preferring to rely on the great expertise of its practitioners. [Yet] the very people who have provided us with wonderful tools for information exchange and processing – word processing, hypertext, the internet – are likely to overtake us and replace linguistic models with informational ones. This threat affects not just language teaching, but all the applications of language knowledge, such as translation, information retrieval, document classification, and lexicography. (Sinclair 1999).

So the revolution is on its way, ready or not. The burning question will be how far we can harness it to help the enterprises of EFL along toward the applicatory adequacy this new millennium imperiously demands.

References

ASTON, G. Corpora in language pedagogy: matching theory and practice. In: COOK, G. SEIDLHOFER, B. (Ed.). *Principle and practice in applied linguistics* Oxford: Oxford University
Press, 1995. p. 257-270.

BARLOW, M. Corpora for theory and practice. *International Journal of Corpus Linguistics* v. 1, n. 1, p. 1-37, 1996.

BEAUGRANDE, R. de. *Linguistic theory: The discourse of fundamental works*. London: Longman, 1991.

BEAUGRANDE, R. de. The 'pragmatics' of doing language science: The 'warrant' for large-corpus linguistics. *Journal of Pragmatics,* v. 25, p. 503-535. 1996.

BEAUGRANDE, R. de. Performative speech acts in linguistic theory: The rationality of Noam Chomsky. *Journal of Pragmatics* v. 29, p. 1-39. 1998.

BEAUGRANDE, R. de. Sentence first, verdict afterwards: On the long career of the sentence. *Word* v. 50, p. 1-31. 1999.

BIBER, D. An analytic framework for register studies. In: BIBER, D., FINEGAN, E. (Ed.).
Sociolinguistic perspectives on register. Oxford: Oxford University Press, 1994.

BOTLEY, S.; GLASS, J.; MCENERY, T.; WILSON, A. (Ed.). *Proceedings of TEACHING and language corpora 1996*. Lancaster: UCREL Technical Papers Volume 9, 1996.

BURNARD, L.; MCENERY, T. (Ed.). *Rethinking language pedagogy from a corpus perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt: P. Lang, 2000.

CARNE, C. Corpora, genre analysis and dissertation writing. In: BOTLEY, S.; GLASS, J.; MCENERY, T.; WILSON, A. (Ed.) *Proceedings of TEACHING and language corpora 1996*.

Lancaster: UCREL Technical Papers Volume 9, 1996. p.127-137,

CHOMSKY N. *Current issues in linguistic theory*. The Hague: Mouton, 1964.

CHOMSKY, N. *Aspects of the theory of syntax*. Cambridge: MIT Press, 1965.

CHOMSKY, N. On binding. *Linguistic Inquiry* v. 11, p. 1-46, 1980.

GHADESSY, M.; HENRY, A.; ROSEBERY, R. (Ed.) *Using small corpora in ELT: theory and practice*. Amsterdam: Benjamins, 2001.

GRANGER, S. (Ed.) *Learner English on computer*: London: Longman, 1998.

HURFORD, J.; HEASLEY, B. *Semantics: A coursebook*. Cambridge: Cambridge University Press, 1996.

JACKSON, H. Corpus and concordance: Finding out about style. In: WICHMANN, A.; FLIGELSTONE, S.; MCENERY, T.; KNOWLES, G. (Ed.). *Teaching and language corpora*. London:
Longman, 1997. p. 224-239.

JOHNS, T. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In: ODLIN, T (Ed.). *Approaches to pedagogic grammar*. Cambridge: Cambridge University Press, 1994. p. 293-313.

KOWITZ, J.; CARROLL, D. Using computer concordances for literary analysis in the classroom. In: JOHNS, T.; KING P. (Ed.). *Classroom concordancing*. Birmingham: Birmingham University English Language Research Journal, Volume 4, 1991, p.135-149.

KUHN, T.S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1970.

KURATH, H. *A word geography of the Eastern United States*. Ann Arbor: University of Michigan Press, 1949.

LEECH, G. Teaching and language corpora: A convergence. In: WICHMANN, A.; FLIGELSTONE, S.; MCENERY, T.; KNOWLES, G. (Ed.). *Teaching and language corpora*. London: Longman, 1997. p.1-24.

LEWANDOWSKA-TOMASZCZYK, B.; J. MELIA (Ed.) *PALC '99: Practical applications in language corpora*. Frankfurt: P. Lang, 2000.

LOUW, B. The role of corpora in critical literary appreciation. In: WICHMANN, A.; FLIGELSTONE, S.; MCENERY, T.; KNOWLES, G. (Ed.). *Teaching and language corpora*. London: Longman, 1997, p.240-251.

MCENERY, T.; WILSON, A. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 1997.

MILTON, J. Lexical thickets and electronic gateways: Making text accessible by novice writers. In: CANDLIN, C.; HYLAND, K. (Ed.). *Writing: Texts, processes and practices*. London: Longman, 1999, p.221-243.

PRATOR, C. The British heresy in TESL. In: FISHMAN, J.; FERGUSON, C.; DAS GUPTA, J.(Ed.). *Language problems in developing nations* . New York: Wiley, 1968, p. 459-476.

QUIRK, R.; GREENBAUM, S.; LEECH, G.; SVARTVIK, J. *A contemporary grammar of English*. London: Longman, 1972.

QUIRK, R..; GREENBAUM, S.; LEECH, G.; SVARTVIK, J. *A comprehensive grammar of the English Language*. London: Longman, 1985.

SAPIR, E. *The fundamental elements of Northern Yana*. Berkeley: University of California Press. 1922.

SINCLAIR, J. Shared knowledge. In: ALATIS, J. (Ed.), *Georgetown University Round Table on Languages and Linguistics 1991*. Washington, DC: Georgetown University Press, 1991, p. 489-500.

SINCLAIR, J.McH. *Corpus, concordance and collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J.McH. *New roles for language centres: The mayonnaise problem*. Pescia: Tuscan Word Centre, 1999.

STEINTHAL, H. *Charakteristik der hauptsächlichen typen des sprachbaues*. Berlin: F. Dümmler, 1860.

TOGNINI BONELLI E. Interpretive nodes in discourse. In: BAKER, M. et al. (Ed.) *Techniques of Description*. Amsterdam: Benjamins, 1993.

TOGNINI BONELLI, E. *Corpus theory and practice*. Pescia: Tuscan Word Centre, 1996.

WENCKER, G. *Sprachatlas des deutschen Reichs*. Berlin: Preussisches Innenministerium, 1887-95.

WICHMANN, A.; FLIGELSTONE, S.; MCENERY, T.; KNOWLES, G. (Ed.). *Teaching and language corpora*. London: Longman, 1997.

WIDDOWSON, H.G. The description and prescription of language. In: ALATIS, J. (Ed.). *Georgetown University Round Table on Languages and Linguistics 1991*. Washington: Georgetown University Press, 1991, p.11-24.

WILLIS, D. Syllabus, corpus and data-driven learning. *IATEFL Annual Conference Report*. 1993. p. 25-31.

WILSON, A.; MCENERY, T. (Ed.). *Corpora in Language Education and Research: A Selection of Papers from TALC94*. Lancaster: UCREL Technical Papers, Volume 4, 1994.