1. Achieving comity: disciplinary border crossings with Guy Aston

We soon learn and soon forget that academic life is inescapably fractured, with different disciplines competing for our allegiance like rival fiefdoms. Part of the painful process of growing up is recognizing which of them provides the most congenial home for us, that is, which discipline best accepts whichever of the many identities we have necessarily been experimenting with throughout our adolescence. When, on meeting a new acquaintance we ask "and what do you do?", we expect to be told not just their hierarchical status (I'm a researcher/ professor/ lecturer/ technician) but also their disciplinary credentials, names which are themselves signs of tribal affiliation. Only a qualified practitioner of (shall we say) "Data-Driven Discourse Analysis" or "English for Academic Purposes" or "Computer Assisted Language Learning" can determine whether a random acquaintance is truly of the same persuasion, a dangerous radical, a would-be disciple, or merely a bluffer. But the names are what matter. Deciding what it is that you really want to do, what it is that you really think, cannot be done in isolation, nor can it go unlabelled. You must find a distinctive name in terms of which to define yourself, making clear what you have in common with existing tribes and where you beg to differ: you must declare your allegiance -- or lack of it.

This is why it is difficult for truly interdisciplinary studies to gain a foothold in the academy, and also why, when they do, they risk succumbing to fossilisation as a new entrenched orthodoxy.

In this paper, I am trying to cross frontiers in various ways. I will use graphics and bullet points in my presentation, as I do when teaching, but the bulk of my argument will be in discursive prose, of the kind I would normally prefer to read silently. I will present some material as if it were part of a course in sociology or history : this is how we used to persuade people to use corpora, or the BNC, or computers more generally, and these are the tools we had at our disposal. How quaint! How naive! But I will also try to show how Guy's attitude to that necessary pedagogy enriched and complemented the activity, and I hope persuade you that his approach has more long-lasting implications and relevance.

I first met Guy in 1965 when we were both making the transition from smart lad at a good school to idiot freshman at an Oxford college. We enjoyed a close and more or less continuous friendship from then until his untimely death last year, experiencing together many of the mundane but essential rituals of growing up – getting a job, finding a place to live, finding a partner, celebrating birthdays and other significant milestones, and so on, and so on. But I am going to talk more about the professional components of our relationship than the personal ones, inextricably intertwingled though the two were. As how could they not be, since, as I have remarked before, Guy made a point of transforming his professional relationships into personal ones wherever possible. I want to focus on the period from about 1992 to 2007, during which our separate academic paths converged, and some effects of that convergence.

In October 1992 I spoke at a workshop in Parma, at the invitation of John Morley though actually (I suspect) instigated by Guy. The gig was to talk about the wonderful new world of "Electronic Resources for Textual Studies" and to persuade a skeptical audience just how much fun they could be having with their computers, beyond simply writing essays on them. (here I am doing it) I recently found a write up I subsequently contributed to "Textus" from which I quote: '... you can use electronic texts to substantiate your subjective impressions about the nature of the language. In teaching, they provide an enormous wealth of information about usage: there is nothing as good as an example(except two examples). And electronic texts are also invaluable as the raw material of other projects, in the creation of multimedia teaching tools with which students can interact, or simply as a basis for cheap publication. For all of this to be possible, of course, there is a need for standardization -- but that is another lecture.'

Amongst many other things, most of which have long since vanished, this article does mention the forthcoming availability of something called the British National Corpus, but the occasion was chiefly memorable to me because it was over an excellent dinner (almost certainly featuring varieties of ham) that I was made quite clearly to understand that all this talk of a global electronic village was all very well, but there was a professionalism in academia, that these were serious people, and you needed a doctorate to express an opinion about how language worked in their company. Fair enough, in retrospect: I was definitely guilty of a bumptious amateurism, a necessary consequence of my evangelical desire to convert the world to SGML. And this did not, I must confess, lessen over the next few years. Over the next few years, as head of the Humanities Computing Unit at Oxford I popped up to harangue innocent academics at all sorts of venues – making a pitch for the British National Corpus, for the Text Encoding Initiative, for something mysterious called Digital Resources in the Humanities, and so on and so on. My carbon footprint (already inflated during the production of the TEI) became seriously overweight; my addiction to the headily addictive work of standards definition showed no sign of diminishing. Here I am, for example, pitching for the TEI at the venerable Haiensa monastery in Korea, in the autumn of 1994. I found I was mutating into a large-ish fish in an admittedly small pond of self-appointed experts in a pleasantly obscure domain: the sort of thing you could guarantee would cause people's eyes to glaze over were they incautious enough to ask "and what do you do exactly?" Fortunately, Guy was around to remind me that the usefulness of a digital resource had a lot more to do with its accessibility than its sophistication.

In April 1994, I was in Lancaster for what we did not then know would become the first of a long lasting biennial series of conferences on Teaching and Language Corpora. I was there mostly to show solidarity with colleagues at Lancaster, who appeared to think there might be some interest in the use of language corpora in language teaching as distinct from language research. To put this in context it is noteworthy that the original design criteria for the

BNC, stated or unstated were predicated on the assumption that a 100 million word corpus of current English, spoken and written, selected according to defined sampling principles, and uniformly encoded would mostly be of interest to NLP researchers and other similarly serious people, beyond of course its core target audience, the professional lexicographers and dictionary makers who had financed it, and the researchers in corpus linguistics who had constructed it. And even this "outreach" as we did not then call it, came about because it was necessary to address the interests of the scientific communities funded by SERC, the Science and Engineering Research Council, in order to secure additional funding from the Department of Trade and Industry. Government funding was also (incidentally) why the regrettable word "National" appears in its title, not being then entirely sullied by association with "nationalism". Follow the money is always a good maxim when trying to understand history. Anyway. Conspicuously missing from the list of target users for the BNC were those who within five years of its release became its primary users: the language teaching community, in particular but not exclusively, those working with non-native speakers of English, whether as teachers, researchers, or learners.

The TaLC conference could be thought of as an attempt to rectify that omission, building as it did on the tradition of "data driven learning" associated with Tim Johns and other pioneers from Birmingham University. Guy gave a rather good paper about how he had been using digitized newspapers as corpora in teaching – CDs containing most of the text of several current newspapers were just becoming available at that time, as were a number of other interesting fore-runners of today's digital collections. But what that first conference demonstrated overall was how much proselytising remained to be done. Although wonderful resources like the BNC were available, the communities which stood most to benefit from their use remained uninformed – when they were not actively hostile to the technical aspects of doing so.

In March 1995, en route to a meeting in Pisa for a deservedly forgotten EU Project called Memoria, I met up with Guy in Florence. My visit report of the time remarks: "I went to Florence, partly for touristic reasons, but mainly to persuade my old friend Dr A to write the BNC Handbook for me, which I duly did over an excellent lunch in an obscure, but very crowded fiaschetteria. This was the sort of Italian eating establishment I like -- hams hanging from the ceiling, elbow-to-elbow diners all shouting at the top of their voices, small children misbehaving with indulgent grandparents, plates of food and bottles of chianti flying everywhere. "

The proposition I put to Guy was to work with me on producing something better than the sort of computer manual which stays on the shelf. Over that excellent lunch we worked out a cunning plan. The book should consist of a series of carefully scripted tutorials, each one focussing on a different and linguistically-interesting question. The tutorials would show with precise instructions how to solve, or at least how to explore solutions for the stated problem using the Windows-based software we had developed at Oxford to search the BNC. (This was called SARA – for SGML Aware Retrieval Application – and entirely by coincidence my second daughter's first name). My part was to help write introductory chapters explaining what the BNC was and what the software did, and also to test drive the tutorials as they took shape. Guy's part was to produce the linguistically interesting questions, to explore how they might be addressed using SARA, and to consider other avenues of enquiry they opened up. Here's a nice photo I found of him doing so.

I also persuaded Guy that we would produce our manuscript in the recently defined TEI SGML format, using descriptive rather than presentational markup, which added a satisfyingly geeky additional layer to the project. It also required me to find ways of generating nicely formatted pages from the SGML input, and required Guy to introduce meaningful SGML tagging into his pellucid prose, like this

For the next couple of years therefore, we grappled together with such classic conundrums as "Do Men Say Mauve"? And "When is ajar not a door?" . Reviewing the text today (you can still find a bootlegged copy of it online if you try hard enough) I am struck by the freshness of the pedagogic approach it takes and the way the reader is treated as a co-conspirator rather than a passive recipient of instruction. Of course, it is now a common orthodoxy to talk up exploratory teaching methods, but I don't know of many instructional manuals from that period which do it so thoroughly. In this respect, the *Handbook* demonstrated clearly Guy's pedagogic principles. Teacher and learner alike seek to achieve comity in a slightly unfamiliar, data-rich, curiously fascinating environment.

Somewhat to our surprise, the BNC Handbook (finally published by Edinburgh University Press in December 1997) sold rather well, at least, until new versions of both the BNC and its software appeared at the end of the 20th century, thus rendering the practical parts of the Handbook entirely useless. We did periodically consider that it might be nice to produce a new updated edition, but the pressure of other work and (let's be honest) a lack of relish for retreading old material kept us from doing so. Even more improbably, the Handbook was translated into Japanese shortly after it became definitively outdated by events.

As I may have already mentioned, my tendencies to evangelism were at their peak at this time. Having endured them himself, Guy hit on the idea of offering me the opportunity to indulge them further by inviting me to run a "TEI training workshop" here in Forli, almost exactly 22 years ago, from 14 to 23 April, 1997, and I am delighted to see that some of those who survived the experience are with us here today, apparently unscarred though perhaps less youthful than they appear in this historic daguerrotype.

This first Forli workshop consisted of eight 90 minute lectures, three two hour practical sessions, and two discussion sessions, somehow squeezed into six days of fairly concentrated effort. It began with lectures on the motivation for encoding and the varieties, benefits, and dangers of markup, before launching into a presentation of what was then known as "document analysis", followed by an exercise in which students were invited to do some. My report of the workshop says drily "The discussion was a little inhibited at this stage, as many of the students were still reeling from the shock of being asked to consider using a computer for something other than word processing."

The reeling may be assumed to have contined the next day, which apparently included "*a whistle-stop tour through the syntax of SGML, requiring considerable amounts of stamina*" before introducing the TEI schema itself and letting the students experience a real-life SGML-aware editor called Author Editor.

Expository lectures on the TEI Architecture, the TEI Header, and a variety of TEI tools, followed, but I am pretty sure that it was the three two hour practicals which made the whole thing bearable. I learned a great deal about how to teach such material from this experience, in particular, the importance of leavening or complementing the technical presentations with ample opportunities for students to experience directly for themselves what the technology could do, whether under the guidance of a scripted exercise, or an open-ended exploration.

This is apparent in the programmes of subsequent TEI workshops (there were 7 in all, in May 98, Dec 99, May 2001, April 2003, June 2004, and March 2006) , which placed a much greater emphasis on practical work, and, crucially, on practical work using as raw material precisely the sort of textual resources which participants were likely to beworking with already, and which they might themselves provide. In other TEI workshops, which tended to attract more diverse groups of specialists, this was not always possible: researchers in mediaeval literature found it needlessly difficult to consider the markup of 19th century postcards, for example. For the language focussed students at Forli, all was grist to the mill, not only the 19th century periodical but also the contemporary web page, the transcribed TV show, the radio interview and so on. Moreover, by 2001, it had become entirely reasonable to require of the participants that they "should have basic computing skills (web browsing, word processing etc.) and a willingness to grapple with alphabet soup" – since that was precisely what they did all the time.

Where the BNC had been designed as a surrogate for the whole of the English language, in some rather ill defined sense, I learned in Forli that there was also a place for smaller corpora which represented very specific kinds of language use. The creation of such "idiolectal" corpora had obvious attractions for student interpreters in particular, since it gave them a way of rapidly identifying lexical or idiomatic patterns characteristic of particular usage domains, such as weather reports, or the antiques trade, or political discussion, or even gothic novels. (I apologise, a little belatedly, for insisting on using *Varney the Vampyre* as my vehicle for showing how to index a corpus with Xaira).

My diary shows that Guy and I spent much of the next decade perfecting a double act, in which I would introduce and/or market the latest version of the BNC, and he would demonstrate its exemplary qualities as a teaching aid in the classroom. He would highlight how this approach might do away with any need for problematic "teacher intuition" about language, placing native and non-native speaking teachers on an equal footing, and above all engaging directly with the learner. Find out about the language and the culture for yourself! Form hypotheses and test them! And, perhaps rather cheekily, "be ready to contradict your teacher".

With some trepidation, let me remind you of some of the salient features of this Astonian approach. My first example is an exploration of the word fallback with serendipitous outcomes

My second example demonstrates how Guy was always ready to exploit data complexity, taking full advantage of the markup in later versions of the BNC to explore the wordtend.

We took this message to many different venues. We spoke at the British Association for Applied Linguistics (Southampton in 1995 and Reading 1999); we spoke at TaLC-inspired conferences such as "Practical Applications of Language Corpora" (Lodz in April 1997 and again in 1999) and "Language Learning and Computers" in Chemnitz in February 1998. To celebrate the millenium we took our double act to the "2nd North American Symposium on Corpus Linguistics" (University of Northern Arizona, Flagstaff, April 2000), which also gave us an excuse to visit the Grand Canyon (by train, of course). We organized a session on corpus linguistics at a meeting of ESSE in Helsinki in August 2000; we spoke at ICAME – the big daddy of all corpus linguistics conferences – in Stratford in 2007; we taught a "Methods Network" workshop in Oxford, jointly organized with Ylva Berglund in November 2007 and repeated in February 2008.

And of course we participated in the biennial TaLC conferences, from 1996 to 2010, by which time Guy (at least) had become a rather benign elder statesman figure rather than an *enfant terrible*. If you're interested in the evolution of that conference, there is a document somewhere on the Internet which lists all the locations, the programmes, and even the abstracts (some of them still online; others requiring an excursion to the Wayback Machine); it also provides links to the various edited proceedings volumes, a skim through which yields some possibly interesting conclusions.

From Lancaster in 1996 to Cambridge in 2018, by way of Oxford (1998), Graz (2000), Bologna/Bertinoro (2002) , Granada (2004), Paris (2006), Lisbon (2008), Brno (2010), Warsaw (2012), Lancaster again (2014), and Giessen (2016), there is ample meat here for the kind of meta-analysis that characterizes interdisciplinary activities in the throes of consolidation. The results of that analysis (there is a good example by Alex Boulton in the most recent TaLC) tend to confirm that yes, it is a good idea to use language corpora in language teaching, and point to a broadly similar set of benefits that such usage can bring.

And yet, TaLC practitioners, some of them, also continue to have reservations about the ease with which traditional language teaching methods can accommodate the full corpus-based, learner-centred, comity-seeking approach that Guy adumbrated thirty years ago. The tools are slicker and more professional, the range and scope of available corpora is magnified beyond recognition, the sense of community amongst those practitioners at all levels is stronger than ever, but some things do not change. Like every other academic tribe before them, perhaps, today's corpus lovers have a self-definition that depends on the existence of an opposing unregenerate tribe of old-fashioned obstructionists. Whatever the reason, every meta-analysis points to the need to continue the struggle of opening minds to new (actually not so new now) ways of engaging with the technology.

By which I do not mean simply using it, but doing so in an informed way. Modern computer systems are very good indeed at assuming they know what you want better than you do. Let’s never cease to doubt that, as Guy unquestionably did. Language and linguistic systems are not simple and data that purports to represent them are inherently and correspondingly complex. No single discipline knows all there is to know about it. We should respect that complexity, and seek to achieve comity in our trans-disciplinary investigations.