Who pays the piper calls the tune: changing linguistic goals in the service of industry. A case study.

David Crystal

I used to be a linguist, pure and simple. I would spend my time bathing in a warm bath of phonetics, grammar, vocabulary, and discourse, in as many dialects, styles, and languages as I could get hold of, and reflecting on the origins, evolution, and character of the human language faculty. Then I became an applied linguist, and things became less pure and less simple. Then I became an applied linguist in the service of industry, and things became extremely impure and not simple at all.

Applied linguistics, as I see it, is a service discipline: the application of the theories, methods, and empirical findings of linguistics to the solution (or at least the clarification) of problems in other domains where language is part of their professionalism. Hammering out what those problems are, in such domains as speech therapy, language teaching, or forensic science, involves a long learning curve, as one has to become thoroughly au fait with the goals of the domain before one dares to begin interacting with it. That is why things become 'less simple'. And in the course of engaging with other professionals, the linguistics is inevitably simplified. Applied linguists are always telling linguistic half-truths. That is why things become 'less pure'.

One of the keynote questions of this conference is: 'How can industry help academia prioritize its research in multimedia, multimodal and multilingual contexts?'. The answer is the same as in any other area of applied linguistics, and it can be simply stated: 'By setting the goals which they want linguists to achieve'. The problems arise in extracting from industrialists a clear statement of what those goals are, and coping with the way these goals change as commercial circumstances alter. The biggest problem facing any applied linguist working in this area is that industry keeps changing the goalposts.

This is not a research field of mine. I can illustrate only from my own experience. But I do have considerable personal experience. Since 1986 half of me has, quite literally, been in the service of industry. Let me outline the sequence of events. In 1984 I left the full-time academic world to become - as the Japanese put it, upset at my self-description as 'free-lance' (which they considered fit only for journalists) - an 'independent scholar'. After a year in which I continued as a linguist, in 1986 I was invited to be editor of an encyclopedia project which had been initiated by a joint venture of Cambridge University Press and W & R Chambers. This was an exercise in general reference publishing. My role was to plan the structure of the work, find contributors, edit their contributions, and bring the project to fruition, which I eventually did in 1990 when the first edition of *The Cambridge Encyclopedia* appeared. Eventually, a whole family of general reference books was published, such as *The Cambridge Factfinder* and *The Cambridge Biographical Encyclopedia*, and each appeared in several editions. I brought together a small editorial team at my home in Holyhead, and built an extension onto the house to accommodate them. From 1986

until 1995, CUP paid me a half salary and covered the costs of my team and their equipment. So during that decade I was a half-time servant of the publishing industry.

In 1995 everything changed. Due to policy changes within CUP, a decision was made to divest themselves of my operation, and they sold the entire encyclopedia portfolio to a Dutch IT firm called AND. AND were not so much interested in the encyclopedias as such but in the classification system which underlay them. Each encyclopedia entry had been classified into a number of categories. The entry on *Winston Churchill*, for example, was classified with reference to politics, journalism, art, literature, and so on, reflecting the many aspects of his life. It thus proved possible to find in our database 'all the novelists', or 'all the 19th century novelists', or 'all the 19th century French novelists', and so on. This approach was a novelty at the time. The classification was devised in the late 1980s, well before Tim Berners-Lee introduced the World Wide Web, and a decade before Google came onto the search scene. AND could see the potential of our taxonomy for improving results coming from the search engines which were around at the time, such as Excite, Lycos, and AltaVista. My main role between 1996 and 2001 was to develop the taxonomy to make it work on the Internet. While the categories I had devised for CUP (within literature, sociology, earth sciences, philosophy, and so on) were still relevant, they did not tell the whole story. Indeed, much of the content found on the Internet was missing. You may find this hard to believe, but sex, fast cars, and travelling to Las Vegas did not figure largely in the editorial remit I had received from the oldest press in the world. The taxonomy grew enormously as a result, from some 500 categories to around 1500. So during that five years I was a half-time servant of the IT industry.

In 2001 everything changed. Due to an over-ambitious acquisitions policy, at a time when dot.coms were failing everywhere, AND went into liquidation, and my editorial office was closed down. Determined not to waste what was by then 15 years of work compiling the encyclopedia database, as well as a taxonomy which contained considerable potential for application (and which had been granted UK and US patents), an ex-AND colleague and I decided to set up our own company to continue developing possible products. We called it Crystal Reference Systems, and from 2001 to 2006 we continued to publish encyclopedias - this time for Penguin Books - and developed the software technology which was required in order to put the taxonomy to work. There were several possible areas of application, which I will illustrate in more detail in a moment: search engine assistance, to improve the relevance of search results; e-commerce, to improve the accuracy of online enquiries; advertising, to improve the appropriateness of ad placement on web pages; automatic document classification, to facilitate the retrieval of information in large electronic databases; and Internet security, to monitor sensitive or dangerous online content. We attracted start-up investment which we hoped would enable us to survive long enough to bring these tools successfully into the market-place. So during those five years, as company chairman as well as director of research and development, I was virtually a full-time servant of my own company.

In 2006 everything changed. The contracts we needed did not come quickly enough to enable us to survive on our own. We needed a partner, and in 2006 we were bought up by a European-wide online advertising business called Adpepper Media. The consequences of this move were that our focus then became exclusively directed towards developing our technology in relation to online advertising, and a

whole new set of goals arose. The taxonomy had to grow in fresh directions to meet advertising priorities: categories had to be devised for hundreds of commercial topics, such as refrigerators, BMWs, and credit cards; and a range of new sensitivities had to be explored. In the world of advertising, it is not simply a matter of ensuring that an ad is appropriately placed on a particular page - ads for suncream or cameras on holiday pages, for example. It is also crucial to ensure that an ad is not placed on an inappropriate page - an ad for children's clothing on an adult porn site, for example. This is quite a tricky area to explore, as I shall shortly explain. But it meant that since 2006 I have been a half-time servant of the advertising industry.

You will have seen from this brief outline how the goals shifted with each employment scenario. The academic CUP, with its concerns for intellectual content and internationalism, led to a demand for a set of encyclopedias as far away from coffee-table books as it is possible to get. With the more downmarket Penguin, the encyclopedias changed their character. The coffee-table came a bit nearer. Several features of the CUP approach were cut (such as the cross-referencing system), and a new series of 'pocket' reference books were produced, requiring a very different set of editorial skills. Or again: with the encyclopedias the taxonomy was a means to an end of finding content-related entries. With AND, this was reversed: now the content was the means to an end of developing a more powerful classification system. My job description changed from being an encyclopedist to being a taxonomist. With Adpepper, there was a further change of direction: everything was now driven by the need to increase advertising revenue in as many countries as possible, and if an aspect of the operation was not seen to be profitable, then it would have no future. An early consequence of this policy was that the encyclopedia side of the business collapsed. Penguin stopped publishing encyclopedias in January of this year, blaming Wikipedia and other online resources for a serious falling-off of print sales. A few months later, finding that other encyclopedia contracts were not forthcoming, Adpepper closed down the encyclopedia division. Who pays the piper, calls the tune. And if pipers do not bring in the income, they have no future. The Internet side, by contrast, went from strength to strength, as will be clear below.

It has been a roller-coaster of a ride, therefore, over the past 22 years, and what is interesting, from the viewpoint of the present conference, is the way my close encounters with these various industries has virtually dictated my research priorities in applied linguistics. For the rest of this paper I will illustrate from the IT side of these activities, as it is here that there has been the most dramatic shifting of emphasis.

The first linguistic task presented by AND, in relation to search engines, was semantic in character. It can be simply characterized. If you are interested in, say, the weather, and you type into a search engine the word *depression*, you will get millions of hits. But the first page of results will give you little or nothing about the weather: what you will get is a page of results offering you advice and drugs about how to cure your state of mind. Can linguistics provide 'search engine assistance', as it is called, to improve the relevance of hits in relation to enquiries? An applied field which I called *searchlinguistics* suggested itself, and it preoccupied me for several years. It might be thought that simply increasing the number of search terms will solve the problem: this turns out not to be so. Because of the way search engines typically work, increasing the number of search terms can bring an increased diversity of results. Some relevant results will be included, of course, but they can be

hidden within a welter of irrelevant hits. And in any case, thinking up exactly which search terms produce the best results isn't always an easy matter.

I decided to approach the task from a lexical point of view. Plainly the problem illustrated by 'depression' arises from the polysemic character of that word. In fact it has four main meanings: psychiatric, meteorological, economic, and geological. If it were possible to devise a filter which distinguished these meanings, the problem would be solved. In our prototype scenario, a search-engine user would type in *depression* and up would come a menu which w'ould say 'Which sense of depression do you require?' and the four contexts would appear. The user would click the one desired, a filter would operate, and only the hits related to the chosen category would be allowed to appear. The question accordingly was: how to provide the content for the filter?

The answer was simple, but time-consuming. To continue with the 'depression' example: if an enquirer wanted to see only web pages to do with the weather, then all we had to do was predict which weather words (technically, lexemes) were likely to appear on the page. If *depression* appears on a page which also contains such items as *rain, low pressure,* and *windy,* then the page is likely to be about meteorology rather than psychiatry or economics. Conversely, if *depression* appears on a page which contains such items as *symptoms, illness,* and *Prozac,* then it probably is not going to be about the weather. So the question now is: how many items are there in the English language which are available to users to talk about the weather - or economics - or psychiatry - or geology? If one could predict what all of these are, then the content of the filter (for these categories) would be comprehensively defined.

Comprehensiveness is critical. Take this example from the field of e-commerce. I once went to a retail website and typed into the search box *mobile phones.* The answer came up: 'We have no mobile phones'. I knew they must have, so I kept trying. I typed in *mobile phone,* then typed both singular and plural forms with and without a hyphen. Nothing. I typed in *cell phone,* spaced, hyphenated, and solid, singular and plural. Nothing. Eventually I discovered that the only enquiry the software would accept was the phrase *cellular phones.* Clearly, most people would not have had the patience to continue their enquiry, and a sale would be lost. It is not a difficult linguistic task to anticipate all the linguistic variants which identify the notion of a 'mobile phone', but all these alternatives have to be included if a site is to anticipate all the ways in which it can be interrogated.

There is of course one place where all the lexical items in a language are comprehensively gathered together: a dictionary. So the research task was now clear: we had to work our way through all the content-specific items in a dictionary (i.e. excluding grammatical words such as *the* and *0[*, and semantically 'light' items such as *make* and *get)* and assign each sense of each item to a category in the taxonomy. We used *Chambers 21st Century Dictionary* as our basic text, and supplemented this by specialist works as required and by Internet searches (which provided most of the proper names - brand names, place names, and the like) that would not normally be included in a dictionary. The task involved several linguistic considerations:

- Semantically, the lexicographers had to identify individual lexical items, their senses, and any high-frequency collocations.

- Grammatically, they had to identify compounding alternatives (is it *flowerpot* or *flower pot?)* and all inflectional variations (such as singular and plural of nouns).
- Sociolinguistically and stylistically, they had to identify formal and informal variants (e.g. *television* and *telly) ,* regional differences (chiefly American versus English variants, such as *boot* and *trunk, color* and *colour),* and within-region spelling alternatives (e.g. *judgment* and *judgement).*

Despite all these variables, the number of items in a category is not as large as you might think: for higher-order categories (such as 'motor vehicles') there might be as many as five hundred; for lower-order ones (such as a specific brand of car) there might be less than a dozen. In its current version there are 2776 keyworded categories in the taxonomy, with an average number of items per category of 104.7, and a range from 5 to 514. An additional and especially important aspect was that each item had to be weighted, to indicate its value as an identifier of a category. For example, the word *quarterback* is a high-value identifier because it occurs only in the category of American Football. *Depression* is a medium-level identifier because it turns up in at least four categories, as we have seen. And *country* is a low-level identifier, because it turns up in dozens of categories (such as in all travel domains).

The entire approach, along with the software which drives it, is now called a *sense engine.* It took three years and a team of forty part-time lexicographers to complete the construction of the lexical database - or, I should say, to complete a first pass, for this kind of task is never-ending. New terms are constantly being introduced into a language, and they have to be added to the database - for example, in 2000 the set of lexical items relating to Iraq did not include the phrase *weapons of mass destruction.* This had to be added after 2003. Or, to take a more commercial example, as new models of motor car come on to the market, their names or model designations have to be incorporated. It takes one full-time person to maintain the database in this way.

The industrial goal in all of this can be summed up in one word: *relevance.* The aim was to find a way of distinguishing relevant from irrelevant search results. When Adpepper took over, this criterion became even more important, in the light of such experiences as the following. A few years ago CNN carried a report of a street stabbing in Chicago. The ads down the side of the screen said such things as 'Buy your knives here' and 'Get good quality knives on eBay'. It is clear what had happened. The primitive software employed by the ad-placement company had found the word *knife* a few times and linked this with the only ads in its database which also used the word - namely, cutlery ads. But the effect was not as intended, and caused much embarrassment. A sense-engine approach would never have produced such a mis-assignment. Because it analyses all the words on the page, and assigns all content words to categories, the classifications of the CNN story would have been to do with crime, policing, and safety. The ad database would then have been searched for ads to do with safety measures and crime prevention.

It should be noted that I just said *classifications,* in the plural. A sense engine makes no assumptions in advance as to what a page is going to be about. It is tested against all 3000 categories, to see which ones are relevant. And there are sometimes surprises. There is a natural intuitive tendency to think that what a page is 'about' is

identified by its headline and first paragraph. But read down to the bottom of a page, and other themes come to the fore. Thus, a report on a win by a tennis star at Wimbledon was rightly classified as tennis; but the sense engine also said the page was about cars and dating. Only by reading the whole page did this become clear. After reporting the tennis win, the writer of the article went on to talk about the star's taste in cars and women. Most web pages are in fact polythematic. It is very rare indeed to find a web page which has just a single theme.

The focus on advertising brought a second goal to the fore, which can also be summed up by a single word: *sensitivity*. There are a number of Internet domains which raise problems for advertisers - for example, sites to do with smoking, drinking, gambling, weapons, pornography, and nudity; sites which present extreme views to do with politics or religion; and sites which introduce a great deal of swearing. Most advertisers (other than those which specialise in such areas, of course) do not want their ads to appear on such sites. How can misplacement be avoided? The arrival of Adpepper made this a new and immediate priority. The sense engine, which had previously been used in a positive way, to include as much as possible, now had to be adapted to exclude. The procedure was the same as before. Each of the dangerous categories had to be explored to identify the set of lexical items which characterized them. Conventional dictionaries were of limited value in this respect, as you can imagine: on the whole, pornographic lexicology has not yet been incorporated into the files of the *Oxford English Dictionary*. It was an interesting few weeks for me, I must admit, as I searched porn sites not looking - as I repeatedly had to assure my wife - at the bodies of the hunks that were there but at the words used to describe their bodies. But we linguists are made of strong stuff, and I survived. And the result was a filter (which I called Sitescreen) which can now flag up sensitive sites so that advertisers can avoid them.

The arrival of Adpepper brought a third goal to the fore, which again changed the priorities of our research. It can also be summed up in a single word: *localisation*. It is all very well having a database in English, but what about other languages? Adpepper had branches in twelve European countries, and the need to provide ad relevance there was as strong as in English-speaking countries. The need to translate the database into their languages suddenly became urgent. It was going to be a huge task. Nearly 300,000 items had to be not just translated, but localised. It is possible to make a straightforward translation from English into these languages for something like three-quarters of the vocabulary. The meteorological sense of *depression* in English will neatly equate to a corresponding word in French, German, and so on. But in around a quarter of cases, there is no direct one-to-one translation, partly for linguistic reasons and partly for cultural reasons. Semantic mismatch is a familiar issue in translation theory, summed up by the popular saying 'The French (or whoever) have a word for it'. Cultural mismatch can be illustrated by the task of translating the names of popular cigarette brands or drinks, which vary from country to country, or by the task of finding what the cultural equivalents are for political or minority groups, especially when used in insulting ways: what is the French equivalent of *Paki*, for instance? This is time-consuming and difficult work, and it has taken about three years to complete for the languages initially selected for translation (German, Danish, Dutch and French).

The goals are continually changing. The ultimate advertising goal is to place ads on web pages so that they relate as closely as possible to the content of the page. If

the page is about Britney Spears, then once upon a time it was enough simply to ensure that the ads were about music, rather than about, say, weapons (spears). Then the demand narrowed: the ads had to be about popular music, and not classical music. Then the demand narrowed further: the ads had to be about Britney Spears as such. The most recent demand requires yet a further narrowing: some advertisers only want their Britney Spears ads to be placed on pages which say nice things about her. If a new album is given a bad review, they do not want to be associated with it. The same point applies to commercial goods. A firm like Hotpoint does not want to advertise on a web page or forum which says that Hotpoint washing machines are rubbish. So now there is a new goal, which can be summed up in another single word: *sentiment.* Can one identify the sentiment of a web page? It is indeed possible, but it requires another lexicographic trawl - this time identifying all the words in a language which express positive and negative attitudes. This linguistic task is trickier. Compare: 'Britney Spears' latest album is rubbish' vs 'Britney Spears' latest album is by no means rubbish'. The reversative force of negative words has to be taken into account. And there are several other syntactic considerations, involving word order and the use of intensifiers (such as *very).* An originally lexical exercise now takes on a grammatical dimension, and the research is forced to move in the direction of the kind of issue that has long been a central concern of natural language processing.

The industrial world is always changing the goalposts, as it responds to what is perceived to be the needs of the customers. Within the course of ten years, my industry-inspired research priorities have changed four times, as summarised in my watchwords: relevance, sensitivity, localisation, and sentiment. Nor is this the end of the story. A recent priority in the advertising industry is behavioural profiling. Here the question is no longer 'oo people like Britney Spears?' but 'oo *you*, John Doe, Mary Smith, like Britney Spears?' Is it possible to tell, from an analysis of your blog, or your page on Facebook or wherever, what your interests are to the extent that a highly personalized advertising campaign can be targeted at you? I am not here concerned with the social or ethical issues involved. It is a complex arena. Speaking personally, there are some ads I would be very happy to receive, relating to my particular interests (a new book on *The Third Man,* for instance, which is my favourite film); and there are others which would irritate me enormously, and fall under the category of spam. This is not a linguistic issue. The question for me is: can behavioural profiling be facilitated by linguistic methodology? I do not think so, but I am not sure.

Whether the research demands can be met is hardly ever a matter of academic judgement. Costs are always there in the background. Is there enough money in the system to pay the pipers? Often there isn't. And if an economic downturn comes along, there definitely isn't. For example, in relation to the localisation exercise, it would have been possible to translate the lexical database into a dozen languages, but the cost of hiring and training translators proved to be prohibitive. The company settled for doing the job by degrees, even though, from an academic point of view, it would have made much better sense to hire everyone at the same time and to have everyone simultaneously available to discuss the sorts of problem that come up. That is how we did it for English in the late 1990s. But a decade on, with a recession looming, and with an eye on falling share prices and the need to demonstrate profitability to the market, the extra investment required was simply not available.

Sometimes the factors influencing research priorities, when one works in connection with industry, are totally unexpected. Let me briefly illustrate from

Internet security. Is it possible to identify a dangerous conversation on the Internet - for example, emails between terrorists or fraudsters, or paedophile grooming in chatrooms? Several Internet and mobile phone organizations have expressed their concern and I was asked to address this issue a few years ago. Yes, it is possible. The same sense-engine methodology can be used, but adapted now to handle a dynamic situation - an ongoing conversation. It is not possible to tell, from a single sentence in a chatroom, whether the speaker is an adult masquerading as a child. But it is certainly possible to tell that a dangerous conversation is taking place if one plots a series of 'leading' sentences over time. Once one has identified the salient lexical items (e.g. in a paedophile context, such words as *clothes* and *wearing*), it is possible to develop a lexical filter which can provide a cumulative score. An innocent conversation will score low; a dangerous one will quickly score high. I devised a procedure of this kind, called Chatsafe, as part of our activities in the early 2000s, but it proved impossible to take it forward. Why? Because to test it, I needed access to real paedophile conversations, and apart from a sample or two provided by child protection agencies on the Web, these proved impossible to get hold of. I contacted the police, the Home Office, and others who had tried to research this area, and the message was the same. If I continued to explore this domain without top-level clearance, I risked arrest! But nobody was able to say how I could get such clearance. And I heard horror-stories - such as the external examiner who was interviewed by police simply for reading a PhD thesis on paedophile activity. It also turned out that even to send such a thesis through the post to an examiner was risking a criminal prosecution. Can such research be done at all, I wonder? The need is there and a possible linguistic solution is available. But in the absence of testing, it remains on the shelf.

After all this, it would actually be a relief to return to a world of pure linguistics, where the task is to identify the phonological constraints governing the use of noun suffixes in a tiny language in the north of Brazil. No risk of arrest there, I imagine. But once the industrial world gets hold of you, it does not willingly let you go. And, I have to admit, having spent so many years working within it, I am reluctant to leave it behind. I want to see whether the various projects are successful - and to understand why, if they are not. Applied linguists have the same curiosity as everyone else. It would be nice to know whether these chronicles, as the Bard put it (in Sonnet 106), are or are not of wasted time.