The computer as a tool for the writer: the quest for the "right" word

The computer as a tool for the writer: the quest for the "right" word

Agnes Kukulska-Hulme and Frank Knowles
Aston University

Of all the linguistic strategies, tools and materials which writers have at their disposal when crafting a text, none can be said to equal in power the impact of the individual word. From the most sensational catchword of an article in the popular press, to the most nuanced meanings of a concept debated in an academic thesis, a judicious choice of key words can prejudge the effect of the text's message on its readers. And such is Man's natural fascination with the individual word, that for centuries He has engaged in gathering all specimens for the showcase of the dictionary, classifying and describing, trying to "tie down" their meaning and surface form. A supreme authority for some, the dictionary's teachings have been deliberately countered or ignored by others. Yet till this day, in its many new guises, it is a source on which many writers rely for inspiration as well as accuracy. It used to be a writer's only tool, save pen and paper; now, in a computerised setting, it has to take its place amongst the battery of tools on offer in what is coming to be known as a "writing environment". Yet because of the status of the word, the dictionary retains a major role. Whether it is still recognisable as a dictionary in the traditional sense is, of course, another question.

The fundamental principle of the dictionary remains, in fact, the same, whatever the details of its environment and its implementation: that is, to serve as a repository of word meanings outside text - or their counterparts in cross-language communication - and their "carriers"; yet it must be happily conceded that computerised implementations of dictionaries have, at least in a R&D setting, immeasurably improved the ergonomics of handling what in traditional dictionary mode tend to be, or to appear to be, large and unwieldy volumes for those wishing to consult them in a sophisticated and multi-faceted way. Computerisation is now also beginning to palpably enhance this environment for those whose everyday tasks revolve around documents: their creation, revision, structure and content. It must remain a constant priority for all concerned with words to encourage further quantitative and qualitative improvements to electronic dictionaries. What has been achieved so far has been achieved, first and foremost, by what might be called the electronic conquest of physical logistic constraints and, secondarily, by innovative approaches, only just being consolidated, to the dictionary as an intellectual concept and artefact.

Some solutions are relatively easy to achieve by computer, even if the results of experimentation are merely played back into a traditional dictionary: the avoidance of definitional circularity, direct or via interposed lexical "third parties", in dictionary material; consistency of approach throughout a publication process potentially

involving scores of different people and extending over a long time-span; the provision of citations which do not have their origin solely in lexicographer's introspection. Other goals are more elusive: an optimal way of structuring and presenting, alongside definitions, all the so-called ancillary information which is so important to and for users, such as valency structures or stylistic markers, for instance. For reasons which are difficult to understand in any terms other than the physical constraints on dictionary size imposed by publishers, usually at their accountants' behest, one of the most crucial forms of information - not just in a bilingual context, but also in a L1 setting - is often either missing, stunted or seriously imbalanced: collocational dynamics. The assumption all along has been that such matters really belong exclusively to the world of text structure and are hence beyond the lexicographer's pale. Yet we all know, not least from the personal statistical database of language behaviour we carry around in our heads, that good collocational control is one of the best hall-marks of linguistic excellence. This particular excellence is the skill of not falling prey to the lure of the cliche and of simultaneously eschewing too great a degree of idiosyncraticity. It seemingly demeans the situation somewhat to merely reduce this problem to the standard task of "lexical insertion" into text but that is effectively what it is so often and this is precisely an area in which computers can deliver enormous assistance to writers, both L1 and L2, especially the latter. It is a matter of some embarrassment, arguably, that it was not until merely five years ago that the first widely available dictionary of English collocations was published. That comment should not, incidentally, serve to diminish the merits of similar works - often of limited part-of-speech coverage - published abroad and clearly intended for speakers and writers of English as a second language. Computers, of course, can be of enormous help in providing this type of writer's aid: the elaboration of such electronic compendia is a task which is driven by surface elements of language and is on the syntagmatic axis even if we talk about co-occurrences and colligations as well as collocations proper. It must be stated that the data management problem is of no mean order, computationally speaking, but the resulting comprehensiveness and flexibility justifies the enterprise. Let us not forget, either, that such writer's aids can be configured to particular functional styles or documentary collections. Once set up, such systems deliver precisely what the writer requires: the temporary display on screen of lexical "attractions" capable of catalysing the writer.

Another potentially highly useful facility - in both senses of the term - provided by computers for writers is that of displaying concordancing material according to various types of lexicographical criteria. This facility is admittedly likely to be of greater use to L2 writers than native speakers: in its essence it is simply a method of calling up for any known headword or potential text word a set - of arbitrary size - of concordanced citations which will permit the writer, after browsing, to make a judicious choice of lexical items based on an immediately preceding analysis of their behaviour in context. This pragmatic approach to lexi-computing for writing

purposes has yet to be launched and advocated; there is little doubt about its potential. Measuring the pay-off is, as with all such facilities, a task requiring the professional assistance of psychologists and others with an interest in human factors research but such considerations should not, at this early stage, contrive to impede R&D work on the facilities themselves.

It might have seemed strange to leave to the end of this survey of computerised lexical aids precisely those facilities and software tools which are already well represented in most word-processing systems and which have gained ready acceptance from users. In the first place one must here identify spelling checkers, the sophistication of which has improved markedly with respect to English-language systems. (The problem with regard to other languages, such as French, German etc. is complicated, of course, because of the more extensive lemmatisation which needs to be performed if "straight" listing does not deliver the goods: however, here too good solid progress is being made.) We are still quite some distance from spelling checkers which will routinely and successfully apply any "intelligence" to running text and when that does become the case the plaudits due will probably be scooped in a much wider context.

In the development and marketing of software tools, conceived as adjuncts to WP/DTP systems, it is a natural progression and urge for lexicographers and software engineers to move on, as quickly as possible, to bigger and better dictionaries. At present, given the commercial unavailability of fully computerised multifunctional lexical compendia, this route rapidly leads, via the constellation of "straight" monofunctional dictionaries, collocational and concordancing resources, to thesaurally-oriented lexical tools. It must be understood that the term "thesaurus" is to be construed in the Rogetian sense as a - ordinary? or special? - type of synonym dictionary rather than in the way characteristic of information scientists. Once again the L1 writer enjoys a distinct advantage directly derived from the hand-held equivalents of these software tools: the hand-held synonym dictionary contains precious little structure, hopefully though sufficient for its fundamental purpose which is to trigger associations in the writer's mind which then lead to a clinch in favour of a particular lexical item. The fertile soil on which this poorly understood process takes place does not have its home inside the computer at all. One appealing and somewhat reassuring view of this phenomenon is that the computer dictionary is merely a temporary extension, an accessory of the human intellect, offering a maximally comprehensive database and a superbly efficient information flow to the human being in charge of the whole process. On this basis consulting a machine dictionary is not different in nature from referring to a hand-held dictionary. The moral of this is simple enough: the onus for producing quality documents rests entirely with writers, theirs is the judicious choice of words - the computer's task is merely to present high-grade lexical information for acceptance or rejection. Perhaps, however, the ease with which computer dictionaries can be consulted will

assist writers, and many others using such systems, to overcome inhibitions often bred during years of schooling that dictionaries are only really to be resorted to in cases of personal language deficit!

It is all too easy to assume that computerised dictionaries of major dimensions and facilities, when they become freely available, will solve all problems! This can hardly be so, given the waywardness of humans and their languages. It is possible for computers to systematise, symmetrise, modularise and homogenise existing dictionaries or newly elaborated lexical holdings but it is not possible for them to optimise dictionaries, in the absolutely literal sense of that word. What they can do is diminish to some extent degrees of previous suboptimality.

Nonetheless the real contribution which computers have already made to mainstream lexicography and, hopefully, will soon make to computer-aided writing as well is the concept and reality of the lexical database or LDB. The LDB gets right away from the rigidity of hand-held dictionaries which are, strange though it may sound at first hearing, compromised by the way they are partitioned in to the left-hand side (LHS) and the right-hand side (RHS). The LHS is, of course, the entry mechanism, the only entry mechanism and it is usually organised according to straight-up alphabetisation of the material incorporated in the dictionary. Alphabetisation is a convenience but it is also a major constraint characterised by its lack of "intelligence". The RHS may be - and usually is - a complex set of fields, strictly delineated from each other and not hospitable to information which does not "fit" its designation. The complexity of dictionary RHS's may be arbitrary and is often further characterised, especially in bilingual dictionaries, by run-on entries. The whole point of this description is to establish that the moment a dictionary of the traditional sort has been printed many pathways to useful lexical knowledge have been blocked off. Flexible topography and flexible retrieval are simply not possible. In a LDB, on the other hand, such things are possible: any field in the entry structure can be used as a sort key, either singly or in combination with others. This maximises retrieval possibilities: it is possible, without indulging in pedantry, overkill or mischief, to enumerate a hundred or so ways of systematically retrieving information from LDB's. It is worth remembering too that LDB's may also be multi-media vehicles incorporating TV, live or canned audio, still and movie photography, animation and computer modelling, as appropriate. Within a very few years LDB's of the above power and sophistication are likely to be readily available - either in-house or via networks - to large numbers of users, in offices, schools, laboratories and lounges.

If, however, the work "arena" is that of the office the likelihood is that the writer is involved with documents which contain a good deal of professional terminology. For writers operating within this type of context the concept of a "term bank" is very attractive: term banks are, in fact, growing in number and size and are also

becoming more widely available on a commercial basis. The origins of term banks are generally twofold: the efforts of international, supranational and national standardisation agencies and the desire of major industrial and particularly manufacturing corporations to establish in-house terminology which is commensurate with obvious needs and which, above all, is internally consistent. Three other minor sources of term banks are worth mentioning: those involved in designing and operating information retrieval systems (IRS's), terminologists/"terminographers", and the translating profession. In this last instance, of course - and potentially in the other cases too - the term bank will have a cross-language capability.

There is neither the time nor opportunity here to enter into the details of terminological studies but two points deserve our attention: firstly, terminologists and information scientists are concerned to systematise technical nomenclatures, principally by purging from them all synonymic series, that is, rejecting and even deprecating all members except just one of such a series; secondly, they strive to build concept systems, usually realised as hierarchies systematically partitioned via subset series, such as levels of abstraction, right down the individual term. Systems of this sort are quite easy to implement on a computer once the initial data analysis has been completed. Retrieval may be via the display of the encompassing network in which the term representing the concept queried has its locus. In a bilingual terminological glossary, of course, retrieval would normally be by one of the more orthodox methods: moving from the LHS to the RHS is an easy piece of navigation when all or virtually all the LHS-RHS linkages are one-to-one.

The idea of structuring a set of terms according to levels of abstraction has an important potential payoff in the general LDB: writers often lose their way as a result of problems involving levels of abstraction. Provided a reasonably workable hierarchisation scheme can be elaborated for appropriate subsets of vocabulary - not an easy proviso to satisfy - such a system could be welcome. Would it be straying out of the purely lexical domain into the business of thematically organising text? If so, what harm thereby? Such a system would also be very akin to the desideratum of providing assistance to writers plagued by the problem of moving from idea to word: that is in fact the working principle of all terminological systems - to name or nominate with respect to artefacts and concepts.

The relationship of the dictionary to the process of writing is difficult to define, not least of all because of the lack of agreement amongst researchers as to what the "process of writing" actually is. Not only is "writing" itself a term which designates a multitude of purposes and styles, it is a term which, when combined with "process", generates an almost infinite array of referents. The process of writing is iterative, adaptive, and ultimately idiosyncratic, but this has not stopped writers from turning to the rigid format of the printed dictionary for help. For in spite of its constraints, it corresponds to a need for direct access to spellings and meanings, and

even to word combinations, idioms, and "traditional" semantic fields (synonyms / antonyms /hyponyms). The more exacting user turns to his own ingenuity as he turns the pages on a more complex trail.

Inasmuch as a typical act of writing is an attempt to convey precise meanings in the most appropriate form, and insofar as its product is available for refinement and review, writing could be said to have a greater in-built educational value than speaking. In this perspective, writing is learning, and tools which support writing should support the learning process which accompanies writing. One of the ways they can do this is for their organisation to be based on some of the principles of learning. They can aim to present information in a progressive manner, building on what has gone before, providing comparison, showing discrepancy and contrast, stimulating interest and the motivation to explore - maybe even giving a hint of the unexpected! And let us not allow the label "learning" to mislead: expert writers are in their own way learners, too.

The hard truth about writing is that nobody can formulate our ideas for us. Even the best support tools will only do just that - support. Any writer, with perhaps the exception of the very creative, will welcome tools which help to collect and organise piecemeal ideas, or to revise a completed piece of text, checking for grammatical or stylistic errors. But where a tool's purpose is to support the manufacturing of meaning, it is no simple matter to know how it could be used, especially at the producing stage, the very act of producing text. With conventional dictionaries, we follow a chain from headword to headword, even from "head-phrase to headphrase", hoping to arrive at the one which fits, or making do with second-best. With computerised dictionaries, information can be "exploded", or processed in parallel or "tail end first".

The choice of any lexical item in the course of writing is governed by factors too numerous to mention here. It may serve our purposes, however, to remind ourselves of just a few: the author's own active vocabulary, subjective preference, and personal goals; the text's readership, its function, its longevity, the scope of its subject domain. On the discourse level, the text's internal cohesion and coherence; and in the sentence or phrase, choice may be further influenced by thematic emphasis, the pursuit of metaphor, even the seemingly trivial pursuit of word play!

The exigencies of this highly complex task of encoding have come to be recognised by the makers of dictionaries for production, notably in the area of dictionaries for foreign learners. There, the particular needs of learners have been reflected in detailed specifications of the permissible syntactic combinations and usual semantic collocations of distinct items of the lexicon. Mother tongue writers have much less "spelt out" to them - and more is the pity! When the needs of writers in respect of production are analysed according to type of textual product and whether a native

or a foreign language is involved, the needs of native writers have equal status to those of foreign learners, but when commercial factors come into play, the former are demoted. Yet the needs of the two groups may not be as disparate as is sometimes supposed.

A widespread assumption, largely justified by the print medium of reference works, is that native writers need no more than to check spellings and meanings, and to find different ways of expressing the same notion. If they are composing strictly technical texts, it is acknowledged that they also need a standardised, reliable vocabulary, and a representation of the concept structure of their domain. No-one could deny that these are actual needs, but there are others. And if we were to spell out our message, we could put it in the form of a plea: "Lead us not from word to word, but from idea to word!"

To try to explain this deliberately exaggerated stance, let us look at some examples of how queries whose goal is to find a fitting word may be formulated and pursued, outside of direct headword look-up:

1. Completion of semantic pairs, series or clusters:

> "strategy and (feel sure there is a second element)"
> -> tactics

2. Search for items to extend a metaphor:

> "specimen - showcase... (what else?)"
> - showpiece - exhibit - curio
> - collector - museum - display
>
> "sea - ocean - water... (what else?)"
> - tide - waves - surf
> - surge - storm - nymph
>
> (fixed uses of this metaphor?)
> sea of faces
> sea of troubles
> troubled waters
> calm before the storm
> doldrums

3. Search by definition + implications:

> "determine beforehand"

-> predetermine?
-> prejudge? Maybe.
Is an unfavourable attitude implied?
(need information on usage)

4. Search by elements of meaning + elements of form (morph. or phonol.):

"idea of going against, opposing.
Verb which starts with counter- or contra-, etc."
-> contradict, controvert, counter
-> counteract, contravene

What are the differences in meaning? (need contexts)

5. Search by opposite meaning + part of speech:

"the opposite of paying attention to detail. Adjective."
-> inattentive, careless, lazy

Need something more positive.

-> free-and-easy
-> daring
-> creative

Yes, close to that.

-> inspired. OK.
(NB: chosen word does not correspond exactly to definition)

6. A second abstract notion, similar in form and meaning:

(to verify correct choice, or as a rhetorical device)

"complementarity" -> completeness
"exposition" -> exposure

7. Other notions which may be entailed (and which may be used to develop the text):

"reconcile" -> compromise

"discouragement" -> dejection, resignation
or disregard, indifference

The point should perhaps be made that these examples of relationships between the words which trigger a search and those which are eventually found do not match the relation types which have oft been described between items of the lexicon when these are looked at in isolation from their realisations in text, though there is, of course, some overlap. Significantly, they cut across the division between paradigmatic and syntagmatic relation types. They differ, too, from the semantic representations of texts generated for automatic language processing, because they attempt to capture the interaction between writer and text.

One major problem is that it may not be possible to catalogue all possible types of interaction. We can, however, derive some comfort from the fact that some searches can potentially serve at least a dual purpose. One device which writers often use is to surprise their readers. They need to know a word's accepted meaning, in order to effect a shift; they need to know a word's standard associations, to be able to introduce unexpected ones. Writers often want to break habits of thought, predigested thinking, prejudice. Even deliberate dissonance or pleonasm can be used for such effect - and we are not talking just about literary writing.

The wide variety of search types, only some of which have been illustrated here, explains in part why no single printed reference work could ever meet these needs. A reference tool which provides all this information, and more, may seem quite unwieldy, even in computing terms. On the other hand, not all lexical items would participate in all search types. This, in effect, is why existing printed sources are so under-used: few writers have to hand - or have the time to handle - a thesaurus, a dictionary of synonyms and antonyms, an idiom dictionary, a dictionary of collocations, a reverse dictionary, a picture dictionary, and so forth.

Our examples underline or hint at some other important points: firstly, the nebulous nature of the search for a fitting word; next, the need to know about the in-built value of words (e.g. their emotional charge, physical force implied, situational context of use, ability to conjure up an image, implication of other notions, and so forth); thirdly, the many paths which can be taken to a given destination, a search becoming widened or narrowed along the way; and fourthly, the interactive nature of a search, which at times is best represented by a dialogue. In computing terms, what we have here are some of the ingredients of an expert system, with a natural language interface!

Perhaps what we really need to do is to see the computerised dictionary within the framework of a decision support system for writing. Some of its information will be hard coded - until such time, of course, as systems can learn to expand and adapt

their own knowledge base! Some of it could come from the system's monitoring of the text as it is being produced, with both a backward- and a forward-looking facility. In the backward sense, this might mean keeping a word frequency tally, and a trace on readability or register; in the more difficult forward sense, suggesting ideas related to the words being used in writing, or to words which were the subject of a previous search.

The automatic processing of conventional dictionary data to extract from it the kind of information that might interest us is already well under way in NLP. As an example, definitions can be dismembered to give components of meaning, and to build up a network of relationships between items of the lexicon. Research work in Knowledge Engineering is also yielding results in the kinds of background information that writers are supposed to have, but in reality either lack or need help with calling to mind. This is especially true when one thinks of background knowledge relating to the EFFECTS of words.

We must constantly remember that the software which we - and many others - regard as appropriate for writing packages can only be built as a result of much inspired, yet painstaking research. It is most important to develop prototype systems which have got the linguistics right! There have been far too many catastrophes in the past where this simple and obvious guiding principle was ignored by supposed NLP developers. Even when a reasonable threshold of linguistic performance has been achieved some significant problems still need attending to: how does one freeze and lock up in a package software that must clearly be highly customisable. Are special text corpora going to be commercially available to those end-users who might wish to conduct their own in-house developments. Will they have access to the right utilities to do that? Will software for writers tend to be adaptive? Will it incorporate self-teach modules and transaction monitors for "power users"? For the moment we can do little more than speculate about these topics.

Words not only express ideas, they organise and control discourse, deliberately focusing it in a given direction, making it move forward. This fact seems to have largely eluded those who in the past have produced the kind of software which claims the name of "ideas processor". Putting your ideas in an arborescent structure is all well and good, but only truly suitable for the writing of a report or for some types of didactic exposition - and it does not help you to formulate those ideas. So even for that one function, it is limited, and it is not an all-purpose tool.

Nor have all-purpose tools ever been of much use to anyone in any area of life, unless they are robustly built and have the flexible advantage of multiple or exchangeable parts. It seems essential to know how the function of texts or parts of texts governs the selection of meanings and words, so that tools - or parts of tools

- may be fashioned for specific ends. Until such essential research is done, we will not have "the right tools for the job."

There is, however, a different line of argument, which runs as follows: it is more important to identify the grand strategies of writing, to find the ways in which content is generated from knowledge of subject, purpose and audience, and to describe the optimal ways in which ideas may be arranged. Words are at too detailed a level to be considered in this approach. Take care of the structure, and the words will take care of themselves.

Quite evidently, these views are not actually opposed. They simply reflect a difference in emphasis, but the difficulty lies in knowing how to make use of their complementary nature. What seems to be lacking is the link - the bridge - between the general and the particular, and this shows in the way that software designed for writers is all too frequently failing to bridge that gap.