Language, cohesion and form: Selected papers of Margaret Masterman

Edited,
with an introduction and
commentaries by Yorick Wilks

Preface

This book is a posthumous tribute to Margaret Masterman and the influence of her ideas and life on the development of the processing of language by computers, a part of what would now be called artificial intelligence. During her lifetime she did not publish a book, and this volume is intended to remedy that by reprinting some of her most influential papers, many of which never went beyond research memoranda from the Cambridge Language Research Unit, which she founded and which became a major centre in that field. However, the style in which she wrote, and the originality of the structures she presented as the basis of language processing by machine, now require some commentary and explanation in places if they are to be accessible today, most particularly by relating them to more recent and more widely publicised work where closely related concepts occur.

In this volume, eleven of Margaret Masterman's papers are grouped by topic, and in a general order reflecting their intellectual development. Three are accompanied by a commentary by the editor where this was thought helpful plus a fourth with a commentary by Karen Sparck Jones, which she wrote when reissuing that particular paper and which is used by permission. The themes of the papers recur and some of the commentaries touch on the content of a number of the papers.

The papers present problems of style and notation for the reader: some readers may be deterred by the notation used here and by the complexity of some of the diagrams, but they should not be, since the message of the papers, about the nature of language and

computation, is to a large degree independent these. MMB (as she was known to all her colleagues) put far more into footnotes than would be thought normal today. Some of these I have embedded in the text, on Ryle's principle that anything worth saying in a footnote should be said in the text, others (sometimes containing quotations a page long) I have dropped, along with vast appendices, so as to avoid too much of the text appearing propped up on the stilts of footnotes. MMB was addicted to diagrams of great complexity, some of which have been reproduced here. To ease notational complexity I have in places used "v" and "+" instead of her Boolean meet and join, and she wrote herself that "or" and "and" can cover most of what she wanted. In the case of lattice set operations there should be no confusion with logical disjunction and conjunction. I have resisted the temptation to tidy up the papers too much although, in some places, repetitive material has been deleted and marked by [….]. The papers were in some cases only internal working papers of the C.L.R.U. and not published documents, yet they have her authentic tone and style, and her voice can be heard very clearly in the prose for those who knew it. In her will she requested, much to my surprise, that I produce a book from her papers. It has taken rather longer than I expected, but I hope she would have liked this volume.

MMB would have wanted acknowledgements to be given to the extraordinary range of bodies that supported C.L.R.U.'s work: the US National Science Foundation, the US Office of Naval Research, the US Air Force Office of Scientific Research, the Canadian National Research Council, the British Library, the UK Office of Scientific and Technical Information and the European Commission.

I must thank a large number of people for their reminiscences of and comments on MMB's work, among whom are Dorothy Emmet, Hugh Mellor, Juan Sager, Makoto Nagao, Kyo Kageura, Ted Bastin, Dan Bobrow, Bill Williams, Tom Sharpe, Nick Dobree, Loll Rolling, Karen Sparck Jones, Roger Needham, Martin Kay and Margaret King. I also owe a great debt to Gillian Callaghan and Lucy Lally for help with the text and its processing.

Yorick Wilks

Sheffield

December 2003

Contents

***Starred chapters have following commentaries by the Editor (and by Karen Sparck Jones for chapter 8).**

I. Editor's Introduction

1. A personal memoir: Margaret Masterman (1910-1986)

Margaret Masterman was ahead of her time by some twenty years: many of her beliefs and proposals for language processing by computer have now become part of the common stock of ideas in the artificial intelligence (AI) and machine translation (MT) fields. She was never able to lay adequate claim to them because they were unacceptable when she published them, and so when they were written up later by her students or independently "discovered" by others, there was no trace back to her, especially in these fields where little or nothing over ten years old is ever reread. Part of the problem, though, lay in herself: she wrote too well, which is always suspicious in technological areas. Again, she was a pupil of Wittgenstein, and a proper, if eccentric, part of the whole Cambridge analytical movement in philosophy, which meant that it was always easier and more elegant to dissect someone else's ideas than to set out one's own in a clear way. She therefore found her own critical articles being reprinted (e.g. chapter 13, below) but not the work she really cared about: her theories of language structure and processing.

The core of her beliefs about language processing was that it must reflect the coherence of language, its redundancy as a signal. This idea was a partial inheritance from the old "information theoretic" view of language: for her, it meant that processes analysing language must take into account its repetitive and redundant structures and that a writer goes on saying the same thing again and again in different ways; only if the writer does that can the ambiguities be removed from the signal. This sometimes led her to overemphasise the real and explicit redundancy she would find in

rhythmical and repetitive verse and claim, implausibly, that normal English was just like that if only we could see it right.

This led in later years to the key role she assigned to rhythm, stress, breathgroupings and the boundaries they impose on text and the processes of understanding. To put it crudely, her claim was that languages are the way they are, at least in part, because they are produced by creatures that breathe at fairly regular intervals. It will be obvious why such claims could not even be entertained while Chomsky's views were preeminent in language studies. But she could never give systematic surface criteria by which the breathgroups and stress patterns were to be identified by surface cues, or could be reduced to other criteria such as syntax or morphology, nor would she become involved in the actual physics of voice patterns.

Her views on the importance of semantics in language processing (which, she continued to defend in the high years of Chomskyan syntax between 1951 and 1966) were much influenced by Richens' views on classification and description by means of a language of semantic primitives with its own syntax. These, along with associated claims about semantic pattern matching onto surface text, were developed in actual programs, from which it might be assumed that she was a straightforward believer in the existence of semantic primitives in some Katzian or Schankian sense. Nothing could be further from the truth: for she was far too much a Wittgensteinian sceptic about the ability of any limited sublanguage or logic to take on the role of the whole language. She always argued that semantic primitives would only make sense if there were empirical criteria for their discovery and a theory that allowed for the fact that they, too, would develop exactly the polysemy of any higher or natural

language; and she always emphasised the functional role of primitives in, for example, resolving sense ambiguity and as an interlingua for MT.

She hoped that the escape from the problem of the origin of semantic primitives would lie in either empirical classification procedures operating on actual texts (in the way some now speak of deriving primitives by massive connectionist learning), or by having an adequate formal theory of the structure of thesauri, which she believed to make explicit certain underlying structures of the semantic relations in a natural language: a theory such that "primitives" would emerge naturally as the organizing classification of thesauri. For some years, she and colleagues explored lattice theory as the underlying formal structure of such thesauri.

Two other concerns that went through her intellectual life owe much to the period when Michael Halliday, as the University Lecturer in Chinese at Cambridge, was a colleague at C.L.R.U. She got from him the idea that syntactic theory was fundamentally semantic or pragmatic, in either its categories and their fundamental definition, or in terms of the role of syntax as an organizing principle for semantic information. She was the first AI researcher to be influenced by Halliday, long before Winograd and Mann. Again, she became preoccupied for a considerable period with the nature and function of Chinese ideograms, because she felt they clarified in an empirical way problems that Wittgenstein had wrestled with in his so-called picture-theory-of-truth. This led her to exaggerate the generality of ideogrammatic principles and to seem to hold that English was really rather like Chinese if only seen correctly, with its meaning atoms, highly ambiguous and virtually uninflected. It was a view

that found little or no sympathy in the dominant linguistic or computational currents of the time.

Her main creation, one which endured for twenty years, was the Cambridge Language Research Unit, which grew out of an informal discussion group with a very heterogeneous membership interested in language from philosophical and computational points of view. Subsequently, the attempt to build language processing programs which had a sound philosophical basis was a distinctive feature of the Unit's work. This approach to language processing, and the specific form it took in the use of a thesaurus as the main vehicle for semantic operations, will probably come to be seen as the Unit's major contributions to the field as a whole, and it was Margaret who was primarily responsible for them. Her vision of language processing and its possibilities was remarkable at a time when computers were very rudimentary: indeed much of the C.L.R.U.'s work had to be done on the predecessors of computers, namely Hollerith punched card machines. Equally, Margaret's determination in establishing and maintaining the Unit, with the enormous effort in fund raising that this involved, was very striking: the fact that it could continue for decades, and through periods when public support for such work was hard to come by, is a tribute to Margaret's persistence and charm. It is difficult for us now, in these days of artificial intelligence in the ordinary market place, and very powerful personal computers, to realise how hard it was to get the financial resources needed for language-processing research, and the technical resources to do actual experiments.

Perhaps the best comment on Margaret's initiative in embarking on language processing research, and specifically on machine translation work, comes from a

somewhat unexpected source. Machine translation, after an initial period of high hopes, and some large claims, was cast into outer darkness in 1966 by funding agencies who saw little return for their money. Reviewing twenty five years of artificial intelligence research in his presidential address to the American Association for Artificial Intelligence in 1985, Woody Bledsoe, one of the long-standing leaders of the field, though in areas quite outside language, said of those who attempted machine translation in the fifties and sixties: "They may have failed, but they were right to try; we have learned so much from their attempts to do something so difficult".

What MMB and C.L.R.U. were trying to do was far ahead of its time. Efforts were made to tackle fundamental problems with the computers of the day that had the capacity of a modern digital wrist watch. Despite every kind of problem, the Unit produced numerous publications on language and related subjects, including information retrieval and automatic classification. For over ten years the Unit's presence was strongly felt in the field, always with an emphasis on basic semantic problems of language understanding. Margaret had no time for those who felt that all that needed doing was syntactic parsing, or that complete parsing was necessary before you did anything else. Now that the semantics of language are regarded as a basic part of its understanding by machine, the ideas of C.L.R.U. seem curiously modern.

Margaret's main contribution to the life of C.L.R.U. was in the continual intellectual stimulus she gave to its research, and through this to the larger natural language processing community: she had wide ranging concerns, and lateral ideas, which led

her, for example, to propose the thesaurus as a means of carrying out many distinct language processing tasks, like indexing and translation. Margaret's emphasis on algorithms, and on testing them, was vital for the development of C.L.R.U.'s work on language processing; but her ideas were notable, especially for those who worked with her, not just for their Intellectual qualities, but for their sheer joyousness.

Her colleagues and students will remember her for her inspiration, rather than her written papers: she made questions of philosophy and language processing seem closely related and, above all, desperately important. On their joint solutions hung the solutions of a range of old and serious questions about life and the universe. In this, as so much else, she was a Wittgensteinian but, unlike him, she was optimistic and believed that, with the aid of the digital computer, they could be solved.

She could not only inspire and create, but terrify and destroy: she had something of the dual aspects of Shiva, an analogy she would have appreciated. Even in her seventies, and still funded by European Commission grants, her hair still black because a gypsy had told her forty years before that it would not go grey if she never washed it, she would rise, slowly and massively at the end of someone's lecture, bulky in her big, belted fisherman's pullover, to attack the speaker, who would be quaking if he had any idea what might be coming. The attack often began softly and slowly, dovelike and gentle, gathering speed and roughness as it went. As some readers may remember, there was no knowing where it would lead.

2. Themes in the work of Margaret Masterman

In this introductory chapter I shall seek to reintroduce and then focus the work of Margaret Masterman by enumerating and commenting briefly on a number of themes in her work. Some of these have been successful, in the sense of appearing, usually rediscovered, in some established place in the field of natural language processing, while others, it must be said, appear to have failed, even though they remain highly interesting. This last is a dangerous claim of course, one that can be reversed at any time. There is in my view a third category, of general programmes rather than particular representational methods, about which one can only say that they remain unproven. In spite of their breath, scope and originality it must also be conceded that Margaret Masterman did not have theories to cover all aspects of what would be considered the core issues of computational linguistics today: for example, she had little or nothing to say on what would now be called text theory or pragmatics. Nor did she have any particular reason for ignoring them, other than that she thought the problems that she chose to work on were in some sense the most fundamental.

The order of the themes corresponds broadly to that of the sections of this book: it moves from abstract concepts towards more specific applications of those concepts, from particular forms to language itself, on which those forms imposed the coherence and redundancy that she believed to be at the core of the very idea of language. I shall continue here the affectionate tradition of referring to her as MMB, the initials of her married name Margaret Masterman Braithwaite.

Ideograms

This was an early interest of MMB's (Masterman, 1954 and Chapter 3) that persisted throughout her intellectual life: the notion that ideograms were a fundamental form of language and were of non-arbitrary interpretation. The root of this idea lay in Wittgenstein's interest (1922) in how pictures could communicate: in how the drawing of an arrow could convey movement or pointing and, before that, in his so-called Picture Theory of Truth, where objects could be arranged to express facts. More particularly, she must almost certainly have been influenced by his Notebooks 1914-1916, where he writes “Let us think of hieroglyphic writing in which each word is a representation of what it stands for”.

The connection of all this to ideograms had been noted by I.A. Richards, who was much preoccupied by Chinese, and who developed English Through Pictures (Richards and Gibson, 1952), a highly successful language teaching tool. MMB came to Chinese through Michael Halliday, then a Cambridge University Lecturer in Chinese, and began to use stick-pictures as representations of situations which could also provide a plausible referential underpinning for language: something universal, and outside the world of the language signs themselves, yet which did not fall back on the naive referentialism of those who said that the meanings of words were things or inexpressible concepts.

Frege (new translation, 1960) had tackled this issue long before and created a notation in which propositions had a sense, but could only refer to the true or the false (at which point all differences between them, except truth value, were lost). This reference to situations, that MMB helped keep alive, has found formal expression

again in Barwise and Perry's Situation Semantics (1983). They, too, wanted a central notion of a situation as what an utterance points to, and they too resort to cartoon-like pictures but, unlike MMB, nowhere acknowledge the role of Wittgenstein's Picture Theory of Truth.

It is as hard to capture the future in this field as of any other, and the movement of a (partially) ideogrammatical language like Japanese to centre stage in language processing may yet show the importance of ideograms for grasping the nature of language. But whatever is the case there, MMB's interest remained not only in the differences in the ways occidental and the main oriental language represent the world, but also in the ways those differences reflect or condition basic thought: she liked to quote a phrase of Whitehead's that our logic would have been better based on the Chinese than the Greeks.

Lattices and Fans

Although not a formalist herself, and considered an anti-formalist by many, MMB nevertheless believed passionately in the applicability of mathematical techniques to natural language; without them, she believed, there would be nothing worthy of the name of theory or science. What she was opposed to was the assumption that formal logic, in particular, could be applied directly to natural language, and she would not concede much distinction between that and the methods of Chomsky (1965), a position that has some historical justification.

The two structures from which she hoped for most were lattices and “fans”, a notion she derived from some work of Brouwer (1952). MMB believed lattices (Masterman,

1959a and Chapter 5) to be the underlying structure of thesauri and fans (Masterman, 1957a and Chapter 4), she believed, mapped the spreading out of the new senses of words, indefinitely into the future. She spent some time trying to amalgamate both representations into a single structure. These efforts have not met with much success nor have they been taken up by others, although Zellig Harris did at one time toy with lattices as language structures, and Mellish (1988) has sought to link lattice structures again to Halliday's categories of grammar and semantics.

Another problem is that fans are too simple to capture much: they have no recursive structure. And lattices are so restrictive: once it is conceded that neither words nor things fall neatly under a taxonomic tree structure, it by no means follows that they fall under a graph as restricted as a lattice either. More promising routes have been found through more general applications of the theory of graphs where the constraints on possible structures can be determined empirically rather than a priori.

Thesauri and the use of large scale language resources

MMB believed thirty years ago that constructed entities like dictionaries and thesauri (especially the latter) constituted real resources for computational language processing (Masterman, 1956, 1959b and Chapters 6 and 7, respectively). That was at a time when any computational operations on such entities were often dismissed, by those working in other areas of computational linguistics, as low-grade concordance work. Betty May compacted the whole of Roget's Thesaurus for MMB, from a thousand “heads” to eight-hundred, and had it put onto punched cards. That formed the basis for a range of experiments on Hollerith sorting machines which contributed to Karen Sparck Jones' seminal thesis work Synonymy and Semantic Classification

(1964, 1986). MMB believed that thesauri like Roget were not just fallible human constructs but real resources with some mathematical structure that was also a guide to the structures with which humans process language. She would often refer to "Roget's unconscious" by which she meant that the patterns of cross references, from word to word across the thesaurus, had generalizations and patterns underlying them.

In recent years there has been a revival of interest in computational lexicography that has fulfilled some of MMBs hopes and dreams. It has been driven to some extent by the availability from publishers of machine-readable English Dictionaries, like LDOCE and COBUILD, with their definitions written in a semi-formal way, one that makes it much easier for a computational parser to extract information from them. But the initial work in the current wave was done by Amsler (1980) at Texas using Webster's, an old-fashioned dinosaur of a dictionary. He developed a notion of "tangled hierarchies" which captures the notion MMB promoted to get away from straightforward tree-like hierarchies.

Current centres for such work include Cambridge, Bellcore, IBM-New York, Waterloo, Sheffield and New Mexico, where it has been carried out by a number of techniques, including searching for taxonomic structures, by parsing the English definitions the dictionary entries, and by collocational techniques applied to the word occurrences in the entries themselves. This last normally involves the construction in a computer of very large matrices, as foreseen in the earlier work of Sparck Jones. Those matrices can now be computed effectively with modern machines in a way that was virtually impossible twenty five years ago.

Although dictionaries and thesauri are in some sense inverses of each other, they also differ importantly in that dictionaries are written in words that are themselves sense-ambiguous, except, that is, for those entries in a dictionary which are written as lists of semi-synonyms (as when, for example "gorse" is defined as "furze" and vice-versa). One of the major barriers to the use of machine-readable dictionaries has been the need to resolve those lexical ambiguities as the dictionary itself is parsed, which is to say, transformed by computer into some more formal, tractable, structure. MMB was more concerned with thesauri than dictionaries as practical and intellectual tools, and they do not suffer from the problem in the same way. Words in a thesaurus are also ambiguous items, but their method of placement determines their sense in a clearer way than in a dictionary: the item "crane", for example, appears in a thesaurus in a list of machines, and therefore means a machine at that point and not a bird. The name "machine" at the head of the section can thus straightforwardly determine the sense of items in it. Yarowsky (1992) returned to Roget as a basis for his fundamental work on large-scale word sense discrimination.

However, the last ten years has seen the Princeton WordNet (Miller 1990) take over from dictionaries like LDOCE as the most used linguistic-semantic resource. WordNet is a classic thesaurus, made up from scratch but with a powerful indexing mechanism and a skeletal set of categories and relations replacing the Roget 1000 heads.

The use of interlinguas

MMB was much associated with the use of interlinguas (or universal languages for coding meaning) for MT and meaning representation (Masterman, 1967 and Chapter

9), and her reply to Bar-Hillel's criticism (1953) of their use has been much quoted. The notion of a uniform and universal meaning representation for translating between languages has continued to be a strategy within the field: it had a significant role in AI systems like conceptual dependency (Schank 1975) and preference semantics (Wilks 1973), and is now to be found in recent attempts to use Esperanto as an interlingua for MT.

MMB's own view was heavily influenced by the interlingua NUDE (for naked ideas or the bare essentials of language) first created by R.H. Richens at Cambridge for plant biology: in a revised form it became the interlingua with which C.L.R.U. experimented. NUDE had recursively-constructed bracketed formulas made up from an inventory of semantic primitives, and the formulas expressed the meaning of word senses on English. Karen Sparck Jones worked on making NUDE formulas less informal, and defining the syntactic form of those entries was one of my own earliest efforts, so that a revised form of NUDE became my representational system for some years. In that system some of Richens' more “prepositional” primitives had their function merged with what were later to become case labels, in the sense of Fillmore's Case Grammar (1968) e.g. Richens' TO primitive functioned very much like Fillmore's Destination Case.

However, MMBs attitude to these primitives was very unlike that of other advocates of conceptual primitives or languages of thought: at no point did she suggest, in that way that became fashionable later in Cognitive Science, that the primitive names constituted some sort of language in the mind or brain (Fodor's view, 1975) or that, although they appeared to be English, the primitives like MOVE and DO were

“really” the names of underlying entities that were not in any particular language at all. This kind of naive imperialism of English has been the bane of linguistics for many years, and shows, by contrast, the far greater sophistication of the structuralism that preceded it.

MMB was far too much the Wittgensteinian for any such defence of primitive entities, in this as in other matters: for her, one could make up tiny toy languages to one's heart's content (and NUDE was exactly a toy language of 100 words) but one must never take one's language game totally seriously (linguists forgot this rule). So, for her, NUDE remained a language, with all the features of a natural one like English or French, such as the extensibility of sense already discussed.

That tactic avoided all the problems of how you justify the items and structure of a special interlingual language claimed to be universal, or brain embedded, of course, but produced its own problems such as that of what one has achieved by reducing one natural language to another, albeit a smaller and more regular one. This, of course, is exactly the question to be asked of the group proposing Esperanto as an interlingua for MT.

She would put such questions forcefully to those in C.L.R.U. who showed any sign of actually believing in NUDE as having any special properties over and above those of ordinary languages, a possibility she had herself certainly entertained: this was the technique of permanent cultural revolution within an organization, known to Zen Bhuddists, and later perfected by Mao Tse Tung.

MMB believed that such interlinguas were in need of some form of empirical justification and could not be treated as unprovable and arbitrary assumptions for a system, in the way Katz (1972) had tried to do by arguing by analogy from the role of assumed "axiomatic" entities in physics like photons or neutrons. One weak form of empirical support that was available was the fact that statistics derived from dictionaries showed that the commonest defining words in English dictionaries (exempting “a” and “the” and other such words) corresponded very closely indeed for the first 100 items or so to the primitives of NUDE. But MMB wanted something more structural than this and spent considerable time trying to associate the NUDE elements with the classifying principles of the thesaurus itself, which would then link back to the distributional facts about texts that the thesaurus itself represented. In this, as in other ways, MMB had more intuitive sympathy with the earlier distributional or structural linguistics, like Zelig Harris, than with the more apparently mathematical and symbolic linguistics of Chomsky and his followers.

The centrality of machine translation as a task.

There is no doubt that MT has become in recent years a solvable task, at least for some well-specified needs, sometimes by the use of new representational theories, but more usually by means of better software engineering techniques applied to the old methods. Merely doing that has yielded better results than could have been dreamed of two decades ago.

MMB must be credited with helping to keep belief in MT alive during long years of public scepticism, and above all with the belief that MT was an intellectually challenging and interesting task (Masterman, 1957b, 1961; Chapters 8 and 10,

respectively). I think that is now widely granted, although it was not conceded within artificial intelligence, for example, until relatively recently. There it was still believed that, although language understanding required inference, knowledge of the world and processing of almost arbitrary complexity, MT did not: for it was a task that required only superficial processing of language. I think that almost everyone now concedes that that view is false.

What MMB sought was a compromise system of meaning representation for MT: one that was fundamental to the process of translation, but did not constitute a detailed representation of all the relevant knowledge of the world. She believed there was a level of representation, linguistic if you will, probably vague as well, but which was sufficient for MT and, in that sense, she totally denied the assumption behind Bar-Hillel's (1953) critique of MT, and which was taken up by some artificial intelligence researchers afterwards (though not, of course, the same ones as referred to in the last paragraph), that MT and language understanding in general did require the explicit representation of all world knowledge. This position of hers cannot be separated from her quasi-idealist belief that world knowledge cannot be represented independently of some language, and hence any true distinction between meaning representation and the representation of world knowledge is, ultimately, misconceived (see her discussion of Whorf in Masterman 1961 and Chapter 10). The only dispute can be about the “level” or “grain” of representation that particular acts of translation require.

In later years she became highly critical of the large EUROTRA machine translation project funded by the European Commission, and surprisingly sympathetic to the old-fashioned American MT system SYSTRAN that she had criticised for many years as

naive. This was partly, I think, because she came to see the vital role of dictionaries for practical MT, a matter that was clear in the development of SYSTRAN, but not (at that time at least) in the linguistic theories that drove SYSTRAN. In a 1979 letter to Margaret King, MMB wrote: "My stance hasn't changed that EUROTRA has got to get clear of the TAUM-approach [the French logical paradigm that underlay early EUROTRA work, Ed.], and to have a major revolution over dictionaries. But there is one question nobody ever asks me, How would you feel if EUROTRA was a triumphant success? Answer; absolutely delighted."

Parsing text by semantic methods

A major concern of MMBs was always how to transform, or parse (Masterman, 1968 and Chapter 12) written English into a machine representation for MT. She believed that such a representation should be fundamentally semantic in nature (i.e. based on meaning rather than syntax) and that those semantic structures should be used in the parsing process itself. The latter view was highly original, since virtually no one had ever proposed such a thing—that doctrine is now known as "semantic parsing", and is well-known even if not as fashionable as it was ten years ago——and espousing it certainly set MMB apart from the prevailing syntactic approaches of her time. Some contemporary clarification will be needed in later commentary on this point, since the meaning of the word "semantics" has changed yet again in recent years. Let us simply add here that "semantic" as used by MMB in this connection cannot be equated with either its use in "semantic grammar" (e.g. Burton 1978) to mean parsing by the use of particular word-names as they occur in text (e.g. as in a program that knew what words would probably follow "electrical"), nor with its currently dominant use in formal, logical semantics, to which we shall return in a moment.

One of MMBs main motivations for her view was that natural languages are highly ambiguous as to word sense, and that fact had been systematically ignored in computational language processing. She went further, and this was again influence from Wittgenstein, and held that they were infinitely or indefinitely ambiguous, and that only criteria based on meaning could hope to reduce such usage to any underlying machine-usable notation. This emphasis set her off not only from those advocating syntactic parsing methods but also from any approach to meaning representation based on a formal logic, including any claim to deal with meaning by the use of set-theoretic constructs that never took any serious account of the ambiguity of symbols.

Historically, MMB was vindicated by the growth of semantic parsing techniques during her lifetime and, although syntactic methods have recently recovered the initiative again, one can be pretty sure the pendulum will continue to swing now it is in motion. In recent years, since the work of Montague, there has been an enormous revival of formal philosophical semantics for natural language, in the sense of set- and model-theoretic methods, that ignore exactly those ambiguity aspects of language that MMB thought so important. Indeed for many theorists "semantics" has come to mean just that kind of work, a development MMB abhorred, not because she did not want a philosophical basis for theories of language, on the contrary, but because she did not want that particular one.

Formal semantics approaches have not yet proved very computationally popular or tractable, and the verdict is certainly not available yet for that struggle. It is worth adding that for other languages, particularly Japanese, MT researchers have continued

to use semantic parsing methods, arguing strongly that such methods are essential for an "implicit" language like Japanese where so much meaning and interpretation must be added by the reader and is not directly cued by surface items.

Breath groups, repetition and rhetoric

These were three related notions that preoccupied MMB for much of her last twenty years, but which have not in my view yet proved successful or productive, and certainly not to MT where she long sought to apply them. This line of work began when she met Guberina, the Yugoslav therapist who managed to reclaim profoundly deaf persons. From him, MMB developed a notion she later called the Guberina Hypothesis (Masterman, 1963 and Chapter 11), to the effect that there were strong rhythms underlying language production and understanding (that could be grasped even by the very deaf), and that these gave a clue to language structure itself. From this she developed the notion of a "breath group", corresponding to the chunk of language produced in a single breath, and that there was therefore a phrasing or punctuation in spoken language, one which left vital structural traces in written language too, and could be used to access its content by computer. Much time was spent in her later years designing schemes by which the partitions corresponding to idealised spoken language could be reinserted into written text.

From there MMB added the notion that language, spoken and written, was fundamentally more repetitive than was normally realised, and that the points at which the repetition could be noted or cued was at the junctions of breath groups. This notion was linked later to the figures of traditional Greek rhetoric, in which

highly repetitive forms do indeed occur, and with the claim that the forms of repetition in text could be classified by traditional rhetorical names.

MMB produced an extensive repertoire of language forms, partitioned by breath groups, and with their repetitions marked: a simple standard example would be “John milked the cows/and Mary the goats” which was divided into two breath groups as shown by the slash, at the beginnings and ends of which were items of related semantic type (John/Mary, cows/goats). Traditional forms of language such as hymns, biblical passages and the longer narrative poets were a rich source of examples for her.

The problem with all this was that it required the belief that all text was fundamentally of a ritual, incantatory nature, if only one could see it, and most people could not. The breath group notion rested on no empirical research on breath or breathing, but rather on the observation that language as we know it is the product of creatures that have to breathe, which fact has consequences even for written text. This last is true and widely accepted, but little that is empirical follows from it.

What is agreed by almost all linguists is that spoken language is, in every way, prior to written. Again, there is agreement among some that the phrase is an under-rated unit, and language analysis programs have certainly been built that incorporate a view of language as a loose linear stringing together of phrases, as opposed to deep recursive structures. Some support for that view can be drawn from the classic psychological work (the so-called "click" effect) that shows that sounds heard during listening to text seem to migrate to phrase boundaries. But none of this adds up to any

view that language processing requires, or rests on, the insertion of regular metrical partitions carrying semantic import.

Again, the claims about repetition and rhetoric can be seen as an extension of a more general, and certainly true, claim that language is highly redundant, and that the redundancy of word use allows the ambiguity of word sense meaning to be reduced. Programs have certainly been written to resolve semantic ambiguity by matching structured patterns against phrase-like groups in surface text: my own early work did that (e.g. Wilks 1964, 1965), and it owed much to MMBs work on Semantic Message Detection. However, the partitions within which such patterns were matched were found by much more mundane processes such as keywords, punctuation and the ends of phrases detected syntactically (e.g. a noun phrase endings).

The oddest feature of MMBs breath-group work, stretching as it did over many years was that it referred constantly to breathing, but nothing ever rested on that: partitions were always inserted into text intuitively in a way that, to me at least, corresponded more naturally to the criteria just listed (keywords, punctuation etc.). Finally, of course, it would be overbold to assert that there will never be applications of Greek rhetorical figures to the computer understanding of natural language, but none have as yet emerged, except their explicit and obvious use as forms of expression. However, in all this one must accept that MMB was one of the few writers on language who took it for granted that the fact it was produced directionally was of some fundamental importance. One can see this, from time to time (as in the work of Hausser, 1999) emerge as a key observation requiring structural exploration, but in most theorizing

about language, such as the transformational-generative movement, this is never mentioned.

Metaphor as normal usage

The claim that metaphor is central to the processes of language use is one now widely granted in natural language processing and artificial intelligence, even if there are few systems that know how to deal with the fact computationally, once it is granted. MMB always maintained that position (Masterman, 1961, 1980 and Chapters 10 and 13, respectively), and the recent rise of "metaphor" as an acceptable study within language processing is some tribute to the tenacity with which she held it. For her it followed naturally from the "infinite extensibility" of language use, the majority of which extensions would be, at first at least, metaphorical in nature. It was one of her constant complaints that Chomsky had appropriated the phrase "creativity", by which he meant humans' ability to produce new word strings unused before, while paying no attention, indeed positively deterring study, of aspects of language she considered universal and genuinely creative. Work such as Fass (1988), Carbonell (1982) and Wilks (1978) carried on her view of metaphor explicitly.

MMB would also welcome anecdotal evidence, of the sort to be found in the work of Cassirer, that metaphorical uses of language were in some historical sense original, and not a later accretion. She rejected the view that language originally consisted of simple, unambiguous, Augustinian names of objects, the view parodied by Wittgenstein (1958, 1972) in the opening of the Philosophical Investigations, and preferred the idea of original primitive atoms of wide, vague, unspecific, meaning, which were then both refined to specific referents in use and constantly extended by

metaphor. Here, for MMB was the root not only of metaphor, but also of metaphysics itself, which consisted for her, as for Wittgenstein, of words used outside their hitherto normal realm of application. But, whereas he thought that words were "on holiday" when so used, for her it was part of their everyday work.

Her critical paper of Kuhn's theory of scientific paradigms (Chapter 13) is an attempt to defend the originality of her own husband (Richard Braithwaite) but what she actually does is to deploy the techniques developed in the chapters of this book as tools to investigate scientific metaphor and analogy empirically, using methods drawn from language processing. This was a wholly original idea. Not to surface again until the artificial intelligence work of Thagard (1982).

Information Retrieval, empiricism and computation

A strong strand in C.L.R.U.'s work was information retrieval (IR): Parker-Rhodes and Needham (1959) developed the Theory of Clumps, and Sparck Jones (ibid.) applied this theory to reclassify Roget's thesaurus using its rows as features of the words in them. MMB touches on IR in more than one of the papers in this volume and she could see what almost no one could at that time, and which many in today's empirical linguistics believe obvious, namely that IR and the extraction of content from texts are closely connected. She believed this because she thought that IR would need to take on board structural insights about language and not treat texts as mere "bags of words", and its not yet totally clear which view of that issue will ultimately triumph (see Strazlkowski 1992).

Much of C.L.R.U.'s theoretical IR work could not be tested in the 1960's: large matrices could not be computed on punched card machines and an ICL 1202

computer with 2040 registers on a drum! It is easy to imagine, looking back, that researchers like MMB guessed that computers would expand so rapidly in size and power, so that the supercomputer of ten years ago is now dwarfed by a desktop workstation. But I suspect that is not so and virtually no one could envisage the way that quantitative changes in machine power would transform the quality of what could be done, in that (once) plainly impossible methods in language processing now seem feasible. It is this transformation that makes it all the more striking that MMB's ideas are still of interest and relevance, since so much has fallen by the wayside in the rush of machine power.

The overarching goal: a Wittgensteinian computational linguistics

There is no doubt that MMB wanted her theories of language to lead to some such goal, one that sought the special nature of the coherence that holds language use together, a coherence not captured as yet by conventional logic or linguistics. Such a goal would also be one that drew natural language and metaphysics together in a way undreamed of by linguistic philosophers, and one in which the solution to problems of language would have profound consequences for the understanding of the world and mind itself. And in that last, of course, she differed profoundly from Wittgenstein himself, who believed that that consequence could only be the insight that there were no solutions to such problems, even in principle.

It is also a goal that some would consider self-contradictory, in that any formalism that was proposed to cover the infinite extensibility of natural language would, almost by definition, be inadequate by Wittgenstein's own criteria, and in just the way MMB

considered Chomsky's theories inadequate and his notion of generativity and creativity a trivial parody.

The solution for her lay in a theory that in some way allowed for extensibility of word sense, and also justified ab initio the creation of primitives. This is a paradox, of course, and no one can see how to break out of it at the moment: if initially there were humans with no language at all, not even a primitive or reduced language, then how can their language when it emerges be represented (in the mind or anywhere else) other than by itself. It was this that drove Fodor (1975) to the highly implausible, but logically impeccable, claim that there is a language of thought predating real languages, and containing not primitives but concepts as fully formed as "telephone". This is, of course, the joke of a very clever man, but it is unclear what the alternatives can be, more specifically what an evolutionary computational theory of language can be.

It is this very issue that the later wave of theories labelled "connectionist" (e.g. Sejnowski and Rosenberg, 1986) sought to tackle: how underlying classifiers can emerge spontaneously from data by using no more than association and classification algorithms. MMB would have sympathised with its anti-logicism, but would have found its statistical basis only thin mathematics, and would have not been sympathetic to its anti-symbolic disposition.

It is easier to set down what insights MMB would have wanted to see captured within a Wittgensteinian linguistics than to show what such a theory is in terms of structures and principles. It would include that same ambiguous attitude that Wittgenstein

himself had towards language and its relation to logic: that logic is magnificent, but no guide to language. If anything, the reverse is the case, and logic and reasoning itself can only be understood as a scholarly product of language-users: language itself is always primary. It is not clear to me whether MMB extended that line of argument to mathematics: I think that she had an exaggerated respect for it, one not based on any close acquaintance, and which for her exempted it from that sort of observation, so that she was able to retain her belief that a theory of language must be mathematically, though not logically, based.

Her language-centredness led her to retain a firm belief in a linguistic level of meaning and representation: she shared with all linguists the belief that language understanding could not be reduced, as some artificial intelligence researchers assume, to the representation of knowledge in general, and independent of representational formalisms (a contradiction in terms, of course), and with no special status being accorded to language itself. Indeed, she would have turned the tables on them, as on the logicians, and said that their knowledge representation schemes were based in turn on natural languages, whether they knew it or not.

On the current concern with a unified Cognitive Science, I think her attitude would have been quite different from those who tend to seek the basis of it all in psychology or, ultimately, in brain research. Chomskyans have tended to put their money on the latter, perhaps because the final results (and hence the possible refutations of merely linguistic theories) look so far off. MMB had little time for psychology, considering it largely a restatement of the obvious, and would I think have argued for a metaphysically-rather than psychologically-orientated Cognitive Science. Language

and Metaphysics were, for her, closely intertwined and only they, together, tell us about the nature of mind, reasoning and, ultimately, the world. She would have liked Longuet-Higgins' remark, following Clausewitz, that artificial intelligence is the continuation of metaphysics by other means.