Julia Lavid 2005: *Lenguaje y nuevas tecnologías: nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI.* Madrid: Cátedra. 407 pp.

Christopher S. Butler
University of Wales Swansea
cbutler@telefonica.net

It is a commonplace that much of our daily life in the twenty-first century, from health care to the remote ordering of supermarket shopping, is crucially dependent on the use of computers. Perhaps less often recognised is the central role played by language in the many computer-mediated processes which affect us. Even those of us who study language in our academic lives are often somewhat ill-informed about the intimate relationship between language and the computer which is such an important feature of the information-based society in which we live. The reasons for this are not hard to find: the underlying theoretical and technological issues are highly complex and require a knowledge, not only of linguistics, but also of computation, not to mention at least a nodding acquaintance with other fields, such as engineering and sociology, which also impact on this truly interdisciplinary area.

In this book, Julia Lavid, herself a well-known computational linguist, has set herself the extremely difficult task of providing an introduction to language and the new technologies which will equip the present-day linguist with the tools for understanding and evaluating developments in this fast-moving field. The book consists of seven chapters, each with a final summary and a substantial section giving details of further reading and relevant web pages.

The first chapter sets the scene by discussing the “information age” in which we live, the result of a technological revolution which rivals in importance the agricultural and industrial revolutions which so changed the world in earlier times. Lavid stresses that in order to understand technological change, we must relate it to socio-cultural context. One aspect of this context in Europe is the recognition that education is not confined to one phase of our lives, but must be a lifelong process, if we are to adapt to the necessities of the modern world. This process of lifelong learning is intimately bound up with the new technological advances in accessing information and making use of it in appropriate ways. The more successful of these new computer-based frameworks for education do not simply present old methods in new wrappings; rather, they exploit the new technologies to provide genuinely new modes of learning, tailored to the needs of the user. While warning us that universities are often very conservative in their approach to learning, the author provides examples of European projects which are already bearing fruit. She also emphasises that the ways in which humanities scholars (including, of course, those whose field is language) do their work are being altered in fundamental ways by the availability of computer-based techniques. In all these aspects, the relationship between language and the computer, set in its socio-cultural context, forms an essential basis.

Chapter 2 is an introduction to the role of the computer in linguistic studies. After pointing out the crucial importance of language in our multilingual European environment, Lavid summarises the advantages offered by the new language technologies: gaining access to the information we really need; communicating orally with computers at home, at work, or in the car; interacting with automatic systems on the telephone; learning other languages; finding out more about what is happening in the world; and exerting an influence on our world, as citizens and as consumers. She distinguishes three types of application: those concerned with written language (spelling, grammar and style checkers; electronic dictionaries and thesuari; automatic subtitling; production of indexes, summaries, etc.), those which deal with spoken language (speech recognition and speech synthesis systems), and those concerned with the interconversion of speech and writing.

All of these applications form part of the area known as Computational Linguistics (henceforth CL), which, unlike traditional approaches to linguistics, treats language in terms of the dynamic processes of its production, understanding and storage, rather than merely as a static product. These processes involve knowledge, not only of language itself, but of the world in which it is embedded, and the situations or contexts in which language is produced. CL is seen by the author as forming part of Cognitive Science, and overlapping with Artificial Intelligence, concerned with computational models of human cognition. As the author points out, work in CL has so far tended to be focused on the *emulation* of human processing systems, geared towards producing similar results, rather than the *simulation* of the actual processes involved in the human mind. In its theoretical aspect, CL formalises the knowledge needed for language use, while in its applied aspect it covers language engineering, or language technologies, concerned with the development of systems which enable users to communicate with computers (database querying systems, information retrieval from documents, man-machine interfaces), to communicate in different languages (machine translation in its various forms, computer-assisted language learning), and to undertake various linguistically-oriented tasks (writing tools, text/corpus analysis programs, lexicographical databases). These technologies can provide new types of employment for graduates with appropriate knowledge and training, and also new opportunities for research and cooperation. The author gives a summary of European and, more narrowly, Spanish initiatives in this area.

From chapter 3 onwards, we turn to the detail of CL, beginning, in this chapter, with computer systems for the understanding of human language. These require two kinds of technology: natural language understanding systems and, for spoken language, also speech recognition programs. The author chooses not to discuss the latter type of application in any detail, claiming that it does not involve true language understanding, but merely assigns a sequence of words to the incoming sound stream. This is, however, surely part of the sequence of procedures needed to achieve an interpretation of the speech input.

The main focus is on the process of *parsing*, that is the assignment of syntactic structures to strings in the input stream. After introducing the basic knowledge needed to understand syntactic analysis, and the need for both a grammar and a lexicon in parsing, Lavid presents an account of the various types of formal grammars proposed in the so-called Chomsky hierarchy (type-0 or unrestricted, type-1 or context-sensitive, type-2 or context free, and type-3 or regular (finite state) grammars), and their advantages and disadvantages for application to the parsing process. She goes on to distinguish three dimensions by which the techniques used in parsing can be classified. Firstly, where alternative structures are possible, processing can follow up these possibilities in *parallel*, or one at a time (i.e. *sequentially*); secondly, the processing can be *top-down*, starting with the sentence symbol and expanding it progressively, or *bottom-up*, starting with the terminal lexical items and trying to combine them into grammatical units; thirdly, parsing can be *non-deterministic*, trying available routes until one or more correct solutions is found, or *deterministic*, i.e. attempting to find the correct solution first time. Systems using some of the most useful combinations of these parameter values are then described. We are also given information on the use of charts to store already parsed fragments, and on the problems caused by ambiguities in the input. New parsing techniques involving the use of probabilistic grammars are also mentioned.

The assignment of syntactic structure is not, of course, enough by itself, but must be supplemented by semantic, lexical and pragmatic analysis. Lavid first presents an account of *ontologies* (i.e. structured lists of concepts and their relationships) from philosophical, cognitive and linguistic viewpoints. This is followed by a discussion of dictionaries and computational lexicons for use in parsing. We then learn about the systems for interpreting pragmatic/discoursal information from the input: anaphor resolution for establishing reference, and the use of *frames*, *scripts* and *plans* to model the organisation of human knowledge. Finally, Lavid discusses the interaction of the various components, dealing with different types of interaction of syntactic, semantic and pragmatic analysis during parsing.

Chapter 4, "El ordenador parlante" ("The Speaking Computer") deals with the generation of natural language. To the question of whether we will get to the point where computers can speak and write like a human being, Lavid's answer is a qualified "yes." Already systems are available which can produce coherent text in restricted domains, such as weather reports, or replies to business orders. The input to a generation system consists of a knowledge base (e.g. meteorological measurements, for a weather report), information on the communicative aim (e.g. summarising data, informing, persuading, narrating), a model of the user (e.g. lay or expert), and a model of the discourse produced up to any particular point. The author discusses the principal components of the generation system itself. *Macroplanning* answers the questions "What information is to be conveyed, and how is it to be structured?" Knowledge is selected from the knowledge base and converted into a text plan (i.e. groups of messages, and eventually individual messages). *Microplanning* then converts the text plan to a sequence of more detailed specifications for (groups of) sentences. Finally, surface *realisation* processes convert the output of microplanning into the final form of the output. This involves the use of a grammar, which is often one specifically adapted for generation. The author points out that systemic functional grammars, which have paradigmatic relationships at their generative heart, have proved particularly suitable in a number of generation projects.

Methods of generation are then described, ranging from the simple use of "canned" text (e.g. in cash machines), through template-based systems (allowing some variation in the general pattern) and phrase-based approaches, to the most sophisticated systems which specify a list of features for each sentence to be generated. Lavid also describes the development of multilingual generation systems and outlines her own work on contrastive grammars for use in generation. A rather short section of some five pages deals with *text-to-speech* systems, and finally some pointers are given towards new applications to the production of document summaries and to the generation of web pages.

Chapter 5, "El ordenador políglota" ("The Polyglot Computer"), is concerned with machine translation (henceforth MT) and with computer-assisted language learning. Lavid sets the context for her discussion of MT by presenting data showing that in present-day Europe, demand for translation far outstrips the capabilities of currently employed human translators. She also deals with some misconceptions which have grown up around this area through poor understanding of what MT is and how it works.

We are then given an account of the two main approaches to MT: rule-based and empirical. In rule-based MT the input text is analysed using a grammar of the source language, and the *transfer* component then converts the resulting abstract representations to corresponding structures of the target language, which are used to produce the output text. Ideally, the transfer step could be dispensed with, if we could devise an *interlingua* which would be abstract enough to represent language-independent underlying structures. However, the author observes that attempts to construct such an intermediate language have shown just how difficult the task is, so that this method is uncommon in commercial systems. She then goes on to discuss empirical approaches to MT: *example-based* approaches, in which aligned texts in two languages are used for the extraction of syntagms which can act as templates for further translations; *translation memories*, storing actual human translations; and *statistical* approaches based on the probability, again calculated from existing translations, that a particular structure in one language will be translated by a specific structure from the second language.

Full, completely automated translation of unrestricted input text remains elusive, not only because of the sheer magnitude of the task and of the resources required, but also due to problems of lexical and syntactic ambiguity, differences in lexical and syntactic structuring between languages, and the difficulty of translating idiomatic expressions and collocations. Lavid explains that current systems have rather more limited goals, and are of two kinds: systems which produce approximate or draft translations, usually in restricted domains, which can then be revised by human translators; and automated tools (e.g. terminological databases, online dictionaries and thesuari, translation memories) which aid human translators in their work and make possible considerable savings in time. In discussing the current limitations of MT, Lavid's account is rather stronger on the problems themselves than on possible ways of overcoming them.

The author's treatment of the use of computers in language learning distinguishes between *authoring* programs, designed to help educators without sophisticated computing skills to construct their own materials, and *user* programs intended for language learners themselves, including exercise and tutorial programs, also tools for more generalised use, such as text processors, graphic design programs, databases, electronic dictionaries, or multimedia encyclopaedias. Lavid particularly emphasises the potential advantages offered by the use of the web in language teaching and learning, highlighting two characteristics, interactivity and student-centred learning, as particularly important. She observes that modern web-based technologies allow users to become part of a virtual community, in which they are not just consumers, but also providers and distributors of information. One minor limitation of the author's account of this whole area is that there is no mention of the feeding of results from computer-based corpus analysis into the provision of the descriptions on which teaching and learning materials can be based. For instance, as early as the 1970s, work was done on the analysis of corpora of German writing in the fields of chemistry and music, in order to limit the range of vocabulary and grammatical structures included in short intensive courses in German for students of these disciplines (see Butler 1974; Grauberg 1981), and that part of the work devoted to vocabulary was carried out using word listing and concordancing software.

Chapter 6 presents an introduction to the usefulness of computers in linguistic research. After emphasising the importance of knowing the limitations, as well as the advantages, of computer-based techniques, Lavid discusses the digital libraries of the future, the availability of bibliographical databases on the internet, and the use of citation indexing facilities on the web. There then follows a section in which the author describes the various stages in the design and planning of a research project. Although the discussion of formulating problems, setting up hypotheses and the identification and classification of variables is of considerable value, the information is not as clearly linked to computational tools as is most of the other information in the book. The author could, for instance, have mentioned mind mapping programs such as FreeMind (currently available for free download from http://freemind.sourceforge.net/wiki/index.php/Main_Page), which help in the generation, organisation and development of ideas, allowing the easy creation and manipulation of maps with nodes representing information of the user's choice, and links between these nodes.

The section on types of corpora and other text collections is a useful overview, in which the author discusses types of corpora, their design and compilation (with appropriate mention of the problems of representativeness), and their annotation to include textual and extra-textual information, also various types of linguistic information (morpho-syntactic, as in corpora tagged for parts of speech; syntactic, in parsed corpora, or "treebanks"; lemmatisation of word forms; prosodic annotation; semantic tagging; incorporation of pragmatic and discoursal information, and software available for these purposes). I would, however, have liked to see here some recognition of the view that any type of annotation potentially distorts the information available in a corpus, by imposing on it classifications of linguistic items at various levels (for discussion see e.g. Tognini-Bonelli 2001: 72–74; Sinclair 2004).

After a lucid summary of the advantages and disadvantages of qualitative and quantitative approaches to linguistic research, Lavid presents an account of quantitative corpus/text analysis which includes discussion of statistical software and word listing and concordancing programs such as WordSmith Tools. The author sensibly warns the reader that although much valuable work can be done with such analytical tools, they may not be suitable for the investigation of complex phenomena which do not have a clear reflex in surface form. The chapter goes on to discuss the usefulness of such facilities for the study of word meaning and use, grammar, discoursal phenomena and language varieties. The details of corpora, software tools and useful web sites at the end of the chapter are quite comprehensive, and clearly an exhaustive listing was out of the question. Among the further resources which, if space had permitted, could usefully have been added are: David Lee's excellent Bookmarks for Corpus-Based Linguists web site at http://devoted.to/corpora; the FLOB (Freiburg-LOB) and FROWN (Freiburg-Brown) Corpora of early 1990s British and American English, respectively (see http://khnt.hit.uib.no/icame/manuals/index.htm); the Corpus of Spoken Professional Academic English (for details see http://www.athel.com/cpsa.html); the national corpora of Eastern European languages such as Czech, Polish and Hungarian; and the concordancers MonoConc Pro and Paraconc (the latter for multilingual analysis), available from http://www.athel.com.

Chapter 7 rounds off the book by discussing the present state of the language industries and predicting future trends, concentrating on three areas: the recovery and extraction of information from the huge quantities of material now available on the web; speech recognition and dialogue systems; and the semantic annotation of web material in order to facilitate the automatic analysis of content.

Julia Lavid's book has three great strengths. Firstly, as I hope to have shown in this review, it offers very comprehensive coverage of a wide range of areas concerned with the use of computers in language studies. Secondly, the author succeeds very well in the extremely challenging task of presenting difficult, highly technical material in such a form that it can be readily understood by those with little or no prior knowledge of the area. Thirdly, it offers, in the sections at the end of each chapter, a wealth of up to date information on sources of further information, both on paper and in electronic form. I recommend the book most warmly to anyone who would like to know more about the fascinating and increasingly important field of computational linguistics.

Works Cited

Butler, Christopher S. 1974: "German for Chemists." *Teaching Languages to Adults for Special Purposes* (CILT Reports and Papers 11). London: CILT. 50–53.

Grauberg, Walter 1981: "Reading Courses in German for Special Purposes." *ADFL Bulletin* 13.2: 24–30.

Sinclair, John M. 2004: "Intuition and Annotation: The Discussion Continues." *Advances in Corpus Linguistics: Papers from the Twenty-Third International Conference on English Language Research on Computerized Corpora (ICAME 23): Göteborg 22-26 May 2002.* Ed. Karin Aijmer and Bengt Altenberg. Amsterdam: Rodopi. 39–59.

Tognini-Bonelli, Elena 2001: *Corpus Linguistics at Work*. Studies in Corpus Linguistics 6. Amsterdam and Philadelphia: John Benjamins.