1. Introduction: Electronic lexicography: From challenge to opportunity

Sylviane Granger

The roots of electronic lexicography, which is only now beginning to be recognized as a field in its own right, date right back to the late 1950s/early 1960s, when it went under the names of computer or computational lexicography and the dictionaries themselves were referred to as machine-readable rather than electronic. At that point, the scope of the field was restricted to “the use of computers in making dictionaries” (Logan 1991: 353). The role of computers at the time appears clearly from Urdang’s (1966) reports on one of the earliest dictionaries to have relied on computer technology, the first edition of the Random House Dictionary of the English Language: “We had been alert to the development of computer technology in the years preceding and we hoped to use computers to do the sorting, codifying, re-arranging, and checking the data at hand, and the text to be written” (ibid, 31). Those expectations were amply fulfilled. However, the computer technology of the time did not extend to the actual production of the dictionary, which was still published by conventional means: “Originally, we had intended preparing a tape for input to some automatic typesetting device. Unfortunately, there were no firms large enough or versatile enough or confident enough of their equipment to convince our manufacturing department that such an undertaking would be feasible” (ibid, 33). In those days, neither lexicographers nor dictionary users had any contact with the computer. Lexicographers “continued to write dictionary entries on paper (increasingly, using structured forms), and it was left to computer specialists to input the data” (Atkins and Rundell 2008, 112). The first dictionary to move from the machine-readable dictionary to the lexical database, thereby greatly helping lexicographers design dictionary
entries, was the *Longman Dictionary of Contemporary English* (1978). Quite a few years elapsed before the first dictionaries began to appear on CD-ROM and dictionary users could truly begin to experience for themselves the benefits of having a dictionary in electronic form. Things really began to accelerate in the 1990s with the rapid development of a range of new mediums, in particular handheld devices and (a little later) online dictionaries. Today lexicography is largely synonymous with electronic lexicography and many specialists predict the disappearance of paper dictionaries in the near future. Symptomatic of this trend is the announcement by Oxford University Press in 2010 to the effect that the next edition of the *Oxford English Dictionary*, the uncontested historical dictionary of the English language, will probably no longer be published in paper format because demand for the online version is so radically outstripping demand for the printed version.

This introductory chapter has three main objectives. First, it aims to define the scope of the field of electronic lexicography as profiled in the volume and give a general overview of the profound changes brought about by the electronic medium. Second, it describes the structure of the volume and briefly summarizes the contribution made by each chapter. And third, it pulls the threads together and sketches some priorities for the future of electronic lexicography.

In this volume electronic lexicography is used as an umbrella term to refer to the design, use and application of electronic dictionaries (EDs), which are in turn defined as primarily human-oriented collections of structured electronic data that give information about the form, meaning and use of words in one or more languages and are stored in a range of devices (PC, internet, mobile devices) (for typologies of EDs, see de Schryver 2003, Nesi 2009 and Lew 2011). However, computer-oriented lexicons, i.e. lexical tools that are primarily designed for use in natural language processing (NLP) applications, are not totally absent from the volume, given that the line between these two types of lexical resources is
progressively narrowing and NLP resources like WordNet are increasingly being integrated into human-oriented tools.

The growing integration of computer technology into dictionaries has led to changes of varying degrees of importance. Overall, these have tended to be rather trivial, often consisting of a mere conversion of the content of the paper dictionary to the electronic medium. Weschler’s (2000) observation that “electronic dictionaries are still fundamentally paper dictionaries on a microchip” is unfortunately still valid for a number of electronic dictionaries today. However, this has started to change and many recent dictionary projects are testimony that the innovations afforded by the electronic medium can radically transform every facet of dictionary design and use. In what follows, I discuss what I believe to be six of the most significant innovations: (1) corpus integration; (2) more and better data; (3) efficiency of access; (4) customisation; (5) hybridisation; and (6) user input. My analysis takes into account both the opportunities they open up and the difficulties they might pose to lexicographers and/or users.

**CORPUS INTEGRATION**

Corpora, i.e. electronic collections of authentic language data, are playing an increasing role in today’s lexicography, both upstream, as raw material which lexicographers mine and refine to produce rich lexical entries, and downstream, as an integral part of the electronic dictionary to which users have direct access and which they can mine for themselves. The role of corpora has become so essential that “no serious compiler would undertake a large dictionary project nowadays without one (and preferably several) at hand” (de Schryver 2003: 167). The abundance and diversity of data provided by corpora coupled with powerful tools designed to handle it quickly and efficiently enable lexicographers to produce much richer descriptions than was the case when the only source of language data came from their own intuition and/or
‘citation slips’. However, the rapid growth of corpora, from collections counted in millions of words to today’s gigantic corpora counted in billions, is leading lexicographers to abandon raw corpus data as primary resource and turn instead to semi-processed data, i.e. data that has already been analysed and sorted into a range of lexicographically meaningful patterns (subject + verb, verb + object, etc.).

MORE AND BETTER DATA

Space limitations have always been one of the main frustrations of dictionary professionals. The change-over to the CD-ROM medium and much more drastically to the internet has lifted these limitations and led to some highly welcome changes like the adoption of a more natural dictionary style devoid of cryptic abbreviations and more importantly, the integration of more and better data: richer collocational coverage, an exponential increase of example sentences, integration of multimedia content (images, graphs, videos, sound files) and extended notes (usage notes, cultural notes, error notes, etc.). However, the old adage ‘too much of a good thing can be a bad thing’ applies here too and great care must be taken not to “swamp the user” (Atkins 1996: 11). In addition, while space restrictions may not be an issue for internet dictionaries, they pose a different but no less acute problem for small-screen devices (smartphones, PDAs, mobile phones), especially in the case of long entries. Highly relevant in this connection is the distinction made by Lew (in press) between ‘storage space’ and ‘presentation space’, the latter referring to the amount of content that can be presented on screen to the user at a given time.

EFFICIENCY OF ACCESS

As rightly observed by de Schryver (2003: 173), “[n]o matter how outstanding the contents of a dictionary, if the contents cannot be accessed in a quick and straightforward way, the
dictionary de facto fails to be a good dictionary”. Multiplicity of access is one of the main benefits of electronic dictionaries. Users have been “liberated from the straitjacket of the printed page and alphabetical order” (Atkins 1996: 2) and now have the choice between a wide range of search options beside the traditional ones (fuzzy search, incremental search, all text search, etc.). In addition, hyperlinks enable them to navigate easily both within and beyond the dictionary. Accessibility is not yet optimal, however, and users still often struggle to get to the right headword or phrase and/or to find the specific information they are looking for. As Lew quite rightly points out (2011: 238), “it is perfectly possible to produce an online dictionary where access is more cumbersome than in a paper book”. Improved accessibility therefore remains a key challenge.

CUSTOMISATION

The amount of data included in electronic dictionaries and the ease with which it can be accessed are worthless if the type of information the user is getting does not correspond to his/her needs. For Atkins (1996: 2), the real challenge of the computer age is “to look at the needs of dictionary users, of every language, and every walk of life, users as diverse as people themselves, and give them the kind of information they need for whatever they are using the dictionary for”. The key word here is ‘customisation’, a process which allows the dictionary to be adapted to users’ needs. Gamper and Knapp (2002) distinguish between two types of customisation: dictionaries can be ‘adaptable’, i.e. involve manual customization by the user, or ‘adaptive’, i.e. adapt automatically to the user thanks to the dictionary logs (for an investigation of internet log files, see Bergenholtz and Johnsen 2005). In both cases, electronic dictionaries are no longer static but dynamic tools and in fact, as noted by de Schryver (2003: 163), they “needn’t really be there anymore, meaning that they might only exist at the time of access”, i.e. they are “virtual” (Atkins 1996: 15). Users’ needs have
become a central issue not only for practical lexicography but also for lexicographic theory. The function theory of lexicography, which underlies several chapters in this volume, “takes the users, the user needs and the user situations as the starting point for all lexicographic theory and practice” (Bergenholtz and Tarp 2003: 172).

HYBRIDISATION
One of the most striking results of the electronic revolution in lexicography is that barriers between the different types of language resources – dictionaries, encyclopedias, term banks, lexical databases, vocabulary learning tools, writing aids, translation tools – are breaking down. Hartmann (2005) refers to this trend as ‘hybridisation’, which he defines as the “combination of one or more types of reference work in a single product”. Among the examples he provides are compromise genres like ‘dictionary-cum-grammar’, ‘dictionary-cum-thesaurus’, ‘dictionary-cum-usage guide’, and ‘monolingual-cum-bilingual dictionary’. According to Varantola (2002: 35), the future electronic dictionary will be “an integrated tool or a number of tools in a professional user’s toolbox where it coexists with other language technology products”. One particularly promising development in this connection is the convergence of electronic lexicography and computer-assisted language learning (CALL) in hybrid tools that Abel (2010) refers to as either ‘dictionary-cum-CALL’ or ‘CALL-cum-dictionary’ according to which of the two functionalities plays the central role.

USER INPUT
The invention of wiki technology has brought about one of the most spectacular changes in lexicography, i.e. the integration of collaborative or community-based input. Whereas in the past, lexicography was a field accessible only to language experts, wiki technology now allows any user to create or amend dictionary entries. As rightly observed by Dutton (2011:
“[i]ncreasing numbers of people who would not describe themselves as lexicographers and may not even know the word now contribute to online dictionaries”. One of the main assets of collaborative lexicography, in addition to its sheer manpower advantage, is its capacity to keep up with language change and lexical innovation: “their very functionality allows them to stay up-to-date, literally cataloguing the language as it is used every day” (Penta 2011). The downside, however, is that accuracy is no longer guaranteed and therefore, in the eyes of some, user-generated dictionaries are “of little scientific value” unless mechanisms for quality control are put in place (de Schryver 2003: 160). Not everybody agrees, however. Penta (2011) insists that we are only beginning to see the potential of collaborative lexicography and that dictionary makers need to “plug into the collective” for fear of “being written out of the future of lexicography”.

This volume explores these issues of corpus integration, more and better data, efficiency of access, customisation, hybridisation and user input from a theoretical, methodological and practical viewpoint, in a range of language use situations. The 19 chapters are grouped into 3 sections: ‘Lexicography at a watershed’, ‘Innovative dictionary projects’ and ‘Electronic dictionaries and their users’.

**Section 1** sets the scene. It introduces some of the major changes that computer technology has brought to the theory and practice of lexicography: new theoretical paradigms; the profound reshaping of the role played by all the practitioners involved in the dictionary-making process; the ever-growing role of corpora and corpus-handling tools; enhanced ways of describing and integrating phraseology; the added value of dictionary writing systems. Taking the standpoint of the dictionary editor and publisher, **Michael Rundell** shows that the recent transfer of many lexicographic tasks from humans to computers is both a blessing and a
curse for the dictionary-making business. For example, the relative ease with which gigantic corpora can now be compiled or purchased has reduced the cost of corpus creation but has led to a massive increase in editorial costs as lexicographers need a lot more time to analyse the data. This in turn has triggered the development of new technologies that are progressively changing the lexicographers’ task from that of writing entries from scratch to editing provisional entries generated by the computer. The chapter provides rare insights into the daily running of ambitious lexicographic projects like the recent DANTE project. The corpus angle is pursued by Adam Kilgarriff and Iztok Kosem who focus on the corpus tools used for lexicographic purposes. An overview of commonly-used tools and their basic features is presented, followed by a more detailed description of the features that are targeted mainly at lexicographers. Particular attention is paid to the main features of the Sketch Engine, currently the leading corpus tool for lexicographers. Some of the most innovative features include word sketches, i.e. automatic summaries of words’ grammatical and collocational behaviour, automatic selection of good examples, facilities for exporting corpus information into the dictionary writing system and a range of customisation options. The chapter concludes with a discussion of possible improvements of current tools. Patrick Hanks’ contribution also examines the role of corpora, in particular the radical possibilities that corpus evidence opens up for the treatment of meaning and patterns of use. Such evidence was inaccessible to pre-corpus lexicographers who could only rely on their own intuitions and a bundle of citations. As a result, they tended to focus on words in isolation and failed to capture the phraseological patterns within which words are used. Hanks describes Corpus Pattern Analysis, the approach he has designed to capture these patterns and map meanings onto them, based on close analysis of corpus evidence in the light of prototype theory. The chapter concludes by considering the pros and cons of wiki-based dictionaries like Wiktionary as a possible model for electronic lexicography of the future. Adopting the lexicographer’s viewpoint, Andrea
Abel provides a survey of the software tools currently used in most dictionary projects to support a consistent and structured lexicographic process from the conception phase to the final product. These tools, called dictionary writing systems (DWS), contain a range of innovative features which liberate lexicographers from the most routine tasks, thereby allowing them to concentrate more on their core business. Abel describes the three main components of DWSs – an editing tool, a database and project management tools – and provides useful keys for choosing between off-the-shelf packages and in-house software. The last part of her chapter stresses the need for DWSs to serve as multifunctional databases from which a range of different dictionaries can be produced according to users’ needs. This leads nicely to Sven Tarp’s chapter which is grounded in the function theory of lexicography, a theory in which user needs are key. Tarp argues that lexicography should be considered part and parcel of a broader consultation discipline integrated into information science. Within this framework dictionaries are seen as utility tools designed to be consulted to satisfy specific needs for information. The main advantage of electronic lexicography is its potential to satisfy the needs of the individual user, by providing new forms of access and presentation of the selected data. Tarp suggests three main methods for achieving this individualization and classifies electronic dictionaries accordingly. Section 1 ends with a chapter by Danie Prinsloo which follows on neatly from Tarp with an investigation into the specific needs of dictionary users in a specific context, the South African context. Though focused on South African languages, the chapter raises the more general issue of electronic lexicography for lesser-resourced languages. Taking recent electronic dictionaries of English as a benchmark, Prinsloo provides a survey of electronic dictionaries for both Afrikaans and Bantu languages and finds very little evidence of the sophisticated features displayed by the English dictionaries. This situation is all the more regrettable as these features would in fact be extremely useful, for dealing with the highly complex morphology that characterizes the
Bantu languages. Prinsloo also argues for the design of dynamic, intelligent and adaptive dictionaries to serve users’ encoding needs and highlights the role that user-generated content can play to speed up the compilation of electronic dictionaries for lesser-resourced languages.

**Section 2** focuses on a number of innovative dictionary projects which incorporate recent developments in metalexicography and/or take advantage of the technological advances described in Section 1. The first chapter by Serge Verlinde and Geert Peeters introduces readers to the new types of search functions that can help develop dictionaries that are more responsive to the needs and profile of the user and thereby pave the way for a more user-oriented lexicography. They show how they have used insights from a usability study to customise data access and search facilities in the *Interactive Language Toolbox (ILT)*, a combined dictionary and learning tool for French and Dutch. One distinctive feature of the *ILT* website is that the user can enter full sentences, paragraphs and texts beside the traditional searches for words and word combinations. Another particularly innovative feature is that the interface is adapted according to the following three types of user needs: decoding, translating and encoding. The primacy of user needs is also in evidence in Magali Paquot’s chapter which describes the *Louvain English for Academic Purposes Dictionary (LEAD)*, an integrated dictionary and corpus tool which can be customised according to the user’s mother tongue and discipline. Corpora play a central role in the dictionary. Corpora of academic texts written by native speakers of English were used to select the academic words and identify their typical phraseological patterning, while learner corpus data allowed for the inclusion of tailor-made error warnings. Users can also query corpora directly within the dictionary interface and here, the *LEAD* innovates by giving access to discipline-specific corpora rather than generic corpora, thus allowing users to visualise academic words in contexts close to their specific situations. The chapter by Natalie Kübler and Mojca Pecman also tackles
language for specific purposes (LSP) but this time in a bilingual perspective. They focus more particularly on innovative ways of introducing collocations in a scientific terminological and phraseological database. They describe the ARTES dictionary, an online bilingual LSP dictionary that can be adapted to the needs of three categories of users - translators, learners and professionals - and put forward solutions for including higher order phraseology (semantic preference and semantic prosody). The chapter also highlights the advantages of including student-generated content validated by experts in an LSP dictionary. In a similar vein Marie-Claude L’Homme, Benoît Robichaud and Patrick Leroyer present DiCoInfo, an online trilingual database which provides a complete lexico-semantic picture of computing and Internet-related terms in English, French and Spanish. They provide a detailed description of the structure of dictionary entries and of the system used for encoding collocations based on Mel’čuk’s theory of lexical functions. They also demonstrate the strength of the system for retrieval and translation and describe the mechanisms they have designed to translate Mel’čuk’s formalisms into user-friendly natural language explanations. Toma Tasovac explores the potential and challenges of basing pedagogical, bilingual electronic dictionaries on WordNet, a comprehensive machine-readable lexical database of the English language which is mainly used in natural language processing. After introducing the specifics of WordNet’s complex lexical architecture and the few pedagogical and bilingual lexicography projects that have used it, Tasovac presents the Transpoetika Dictionary, a collaborative, open-access bilingualised Serbian-English learners’ dictionary aligned with WordNet. The author also discusses strategies for expanding the scope of WordNet-based dictionaries through web services and social media platforms such as Flickr and Twitter. The ensuing chapter by Christian M. Meyer and Iryna Gurevych investigates the status and implications of collaborative dictionaries through a comprehensive description of Wiktionary, a freely available, collaborative online lexicon. They study the variety of encoded lexical, semantic,
and cross-lingual knowledge of three different language editions of Wiktionary – English, German and Russian – and compare the coverage of terms, lexemes, word senses, domains, and registers to a number of expert-built lexicons in the three languages. In their conclusion the authors suggest a number of categories of users who might benefit most from Wiktionary. The last chapter in the section, by Jette H. Kristoffersen and Thomas Troelsgård, demonstrates the potential of the electronic medium for sign language dictionaries. The main problem posed by sign languages for lexicographers is that they use visual modality and have no written representation that is commonly used among native signers. The authors start by describing some of the essential differences between dictionaries of sign language and dictionaries of spoken/written language, and describe some of the features that exploit the potential of the electronic medium, in particular the inclusion of video recordings and the possibility of searching on the basis of handshape and place of articulation. They focus on the Danish Sign Language Dictionary and describe some of the functionalities that have been implemented to accommodate the needs of different categories of users.

Section 3 adopts the users’ perspective and investigates the effectiveness of electronic dictionaries for different categories of users, in particular students and translators. The first chapter by Anna Dziemianko compares the respective use and usefulness of paper and electronic dictionaries as evidenced by several recent empirical studies. Various formats of electronic dictionaries are taken into consideration and the following functionalities are investigated: decoding, encoding, speed, look-up frequency, learning and appreciation. The wide variety of studies reviewed raises issues of comparability. However, some interesting trends emerge. For example, while electronic dictionaries prove to greatly reduce search-related lexicographic costs, paper dictionaries appear to be better language learning tools. The chapter ends with possible directions for further investigations into paper vs. electronic
Robert Lew examines two areas in which current electronic dictionaries can be further improved so as to serve human users better: access to lexicographic data and inclusion of novel types of data. In terms of access, the author considers how electronic dictionaries can help in situations where users are unsure about the spelling of the word they want to look up; efficient entry navigation and access to multi-word expressions are also discussed. As regards types of data, Lew discusses the potential benefit of including different categories of multimedia data – audio (recorded or synthesized), static pictorials, animations and videos – and concludes that not all of them may be equally useful for dictionary users.

Hilary Nesi explores some popular types of electronic dictionary that are less prestigious than those produced by established research centres or publishing houses. These dictionaries, which she calls ‘alternative e-dictionaries (AEDs), have been largely ignored in the academic literature. Nesi reviews previous surveys of teacher and student attitudes to AEDs in their various formats: online, on disc, and in pocket electronic dictionaries. She examines the wide range of information found in AEDs. Sometimes they include dictionaries from reputable publishing houses (though the AEDs tend to be vague about their exact provenance), but they typically also include much more dubious lexicographical material of uncertain origin, as well as raw data taken directly from the web. The chapter encourages teachers and metalexicographers to critique AEDs so that users’ choices can be better informed.

Lynne Bowker considers translators’ lexicographic needs, exploring how these might be better met in the age of electronic lexicography. A variety of possibilities, covering both content-related and presentation-related aspects, are considered. These include combining general and specialized content, adding frequency and terminometric data, explicitly identifying relations, providing additional usage information (including cautions against inappropriate usage), integrating more multimedia content, standardizing design of search techniques, introducing customization options, and creating integrated collections of resources. The chapter ends with
a plea for increased communication between lexicographers and translation experts. **Pedro A. Fuertes-Olivera** takes a look at free internet dictionaries and examines their usability for the teaching and learning of Business English in the light of the function theory of lexicography. The analysis of a range of features of free internet dictionaries (links to internet texts and videos, integration in a language portal, updated data, monolingual/bilingual hybridisation) coupled with a comparison of the information displayed in free internet dictionaries and some widely-used printed dictionaries leads Fuertes-Olivera to conclude that, as long as users are trained to use them critically, free internet dictionaries can be reliable teaching and learning tools. The last chapter by **Carolin Müller-Spitzer, Alex Koplenig** and **Antje Töpel** summarizes some key findings of a research project on the use of online dictionaries, using established methods of empirical social research. Through an analysis of the results of two extensive online surveys conducted in 2010 with more than 1,000 participants, they present four different aspects of practical relevance for electronic lexicography: (1) which electronic devices are used for online dictionaries; (2) which type of screen layout is best suited; (3) how users rate different characteristics of online dictionaries; (4) how users form evaluative judgments about innovative features of online dictionaries such as multimedia elements or user-adaptive access. The study shows that when combined with other methods, in particular the analysis of log files, questionnaire surveys can provide highly valuable insights into online dictionary use.

Lexicography is clearly at a turning-point in its history. In the words of Bergenholtz et al. (2009: 8), it “has reached a crossroads where it is difficult to develop further without a thorough rethink”. Developments have never been as quick and diverse, and language needs are growing exponentially. And yet the field of dictionary-making is in danger. Scientific articles, blogs and newspapers resonate with bleak forebodings about the future of
lexicography. For Wallraff (2009) “now that (...) all of us Internet users can find out for
ourselves much of what we do want to know, dictionaries’ days may be numbered”. It is
paradoxical that dictionaries should be under threat precisely at a time when, as rightly
observed by Rundell in this volume, “dictionaries have at last found their ideal platform in the
online medium”. This volume is a direct reflection of the current state of electronic
lexicography. It shows that all the facets of the field are undergoing a transformation so
profound that the resulting tools bear little resemblance to the good old paper dictionary. The
developments are clearly promising but many challenges remain, especially for dictionary
publishers who have to compete with purveyors of free online dictionaries. In this climate it is
essential that users go for the good tools rather than using anything that happens to be out
there on the web. But this will only happen if, a) all the people involved in the dictionary-
making process have the user perspective as their top priority and b) users are properly
trained. The user perspective runs as a unifying thread throughout the volume. Significant
efforts are being made to identify user needs and bring to bear all the available technology to
try and meet them. However, as pointed out by experts and non-experts alike, there is still
ample scope for improvement. Wallraff (2009), for example, deplores the fact that
lexicographers “invest a lot of hard work in things users don’t need or want” and de Schryver
(2003: 163) regrets that “methods to avoid swamping the user are still very much
underdeveloped”. The abundance of language-oriented websites and the wealth of language
descriptions they contain lead to an ‘embarras de richness’ which may well be
counterproductive. This is especially true of online dictionaries where “without proper
guidance users run the risk of getting lost in the riches” (Lew 2011: 248). While dictionary
publishers are certainly wise to invest time and energy in Search Engine Optimization to
enhance their online visibility (cf. Lannoy 2010), it is important that they continue to allocate
a large portion of their resources to another equally important type of optimization, which to
parallel Search Engine Optimization, one could label ‘Dictionary Use Optimization’, entirely focused on tailor-making the dictionary to users’ needs. At the same time, we need to bear in mind that dictionaries have become quite complex tools which require specialized training. Even if increased attention is being paid to users’ specific needs, one cannot take it for granted that dictionaries will be used appropriately. Weschler (2000) reminds us that “[i]t takes time to learn to use the functions of an ED [electronic dictionary]”. As the electronic medium is gradually moving lexicography into the more general field of ‘information science’ (Tarp 2011 and this volume), training could take the form of a lexical reference module integrated into a much-needed electronic literacy programme spanning the whole curriculum. This would ensure that users are equipped with the necessary knowledge to tell the difference between a good tool and a bad one. The new challenges that the field of lexicography is currently facing are so diverse that they call for multi-disciplinary teams involving, besides experts in lexicography, specialists in IT, corpus linguistics, translation and, for the many learner-oriented tools, second language acquisition and computer-aided language pedagogy. With the right priorities and the right teams, I have no doubt that it will be possible to turn the challenges that we face into as many opportunities for success to the benefit of all involved.