Lexicology and Lexicography

Wolfgang Klein, Max Planck Institut für Psycholinguistik, Nijmegen, The Netherlands

Abstract

Lexicology is the scientific investigation of the lexicon of a language, including, for example, its historical development, its social stratification, its quantitative composition, or the way in which some thematic area is encoded. Lexicography, the oldest subdiscipline of linguistics, deals with the compilation of dictionaries. There are many types of dictionaries, depending mainly on which lexical units are included, and which of their properties – such as sound, spelling, grammatical features, meaning, etymology, and others – are described. Two technological innovations are crucial to the presentation of lexical information – the art of printing, which led to the familiar printed dictionary, and the computer, which makes it possible to build up and to exploit large data sources and to develop complex digital lexical systems.

Introduction

Each language has a lexicon and a grammar, i.e., a set of elementary expressions and a set of rules according to which complex expressions are constructed from simpler ones. Some of these rules form complex words; others operate beyond the boundaries of the word, thus producing phrases and sentences. These distinctions, familiar from the days of the Greek grammarians, are not always clear-cut, for at least two reasons. First, the notion of 'word' is not very well defined. Second, there are complex expressions, whose meaning is more or less predictable from the meaning of its components, whereas this is not true for other complex expressions. The former are said to be 'compositional,' whereas the latter are 'lexicalized'; slightly different terms to characterize this opposition are 'productive' vs 'idiomatic,' and 'free' vs 'fixed'; in each case, the distinction is gradual. Lexicalization is rarely observed for inflected words (a possible exception are 'participles' such as *crooked* in *a crooked street*), but very frequent for compound words, such as *landlord* or (*to*) *withdraw*, or phrases such as *to kick the bucket*, which has a compositional as well as a lexicalized reading. Do lexicalized expressions belong to the lexicon of a language or to its grammar? There is no straightforward answer; their form is complex and rule based, their meaning is not. Therefore, it is useful to take the term 'lexicon' in a somewhat broader sense; it contains all elementary expressions (= lexicon in the narrower sense) as well as those expressions which are compound in form but not accordingly in meaning (*see* Lexicon). The scientific investigation of the lexicon in this sense is usually called lexicology; it includes, for example, the historical development of the lexicon, its social stratification, its quantitative composition, or the way in which some subfield is encoded in lexical items (e.g., 'terminology of hunting,' 'verbs of movement'). Lexicography, by contrast, deals with the compilation of dictionaries. There is considerable overlap between both disciplines, and in fact, not all authors make such a terminological distinction.

The Lexicon

The lexicon of a language is stored primarily in the head of its speakers, and for most of the history of mankind, it was only stored there. We do not know what form the 'mental lexicon' has (*see* Psycholinguistics: Overview). There is agreement, however, that it consists of individual lexical units which are somehow interrelated to each other. There is no generally accepted term for lexical units. The familiar term 'word' is both too broad and too narrow; one would not want to consider *goes* as a lexical unit, although it is a word, whereas expressions such as (*to*) *knock out* or *red herring* are lexical units but consist of several words. Other terms occasionally found are 'lexeme,' 'lemma,' or 'lexical entry,' but since these are also used in other ways, it is probably best to speak of lexical units.

It is important to distinguish between a lexical unit and the way in which it is named. The word *house* in a dictionary, followed by all sorts of explanations, is not the lexical unit – it is a name for such a unit. The lexical unit itself is a bundle of various types of properties. These include:

1. phonological properties, which characterize how the lexical unit is pronounced; they include sounds, syllabic structure, lexical accent and, in some languages, lexical tone;
2. graphematic properties, which characterize how the lexical unit is written (*see* Spelling, Psychology of);
3. morphosyntactic properties, which characterize how the unit can become part of more complex expressions; typically, they concern word class, inflectional paradigm, and government relations;
4. semantic properties, which concern the 'lexical meaning' of the unit, i.e., the contribution which it makes to the meaning of the construction in which it occurs.

Some of these properties may be absent. This is most obvious for graphematic properties, since not all languages are written. There are a few lexical units without lexical meaning, such as the expletive *there* in English. Many linguists also stipulate 'zero elements,' i.e., units with morphosyntactic and semantic properties but without phonological properties (such as 'empty pronouns'); but these are normally treated in the grammar rather than in the lexicon.

Whereas these four types of properties are the defining characteristics of a lexical unit, other information may be associated with it, for example, its etymology, its frequency of usage, its semantic counterpart in other languages, or encyclopedic knowledge (thus, it is one thing to know the

meaning of *bread* and a different thing to know various sorts of bread, how it is made, its price, its role in the history of mankind, etc.).

The lexical units of a lexicon are in many ways interrelated. They may share some phonological properties (for example, they may rhyme with each other), they may belong to the same inflectional paradigm, they may have the opposite meaning ('antonyms,' such as *black* and *white*), approximately the same meaning ('synonyms,' such as *to begin* and *to start*), or when complex in form they may follow the same construction pattern. Lexicological research is often oriented toward these interrelations, whereas lexicography tends to give more weight to the lexical unit in itself. In general, there is much more lexicographical than lexicological work (for a survey of the latter, see Cruse et al., 2005); in fact, if there is any piece of linguistic description for some language, it is probably an elementary bilingual dictionary. The depth of this work varies massively not only across languages, but also with respect to the particular lexical properties. Whereas the phonological, graphematic, and morphosyntactic features of the lexicon in Latin, English, French, and some dozen other languages with a comparable research tradition are fairly well described, there is no theoretically and empirically satisfactory analysis of the semantics of the lexicon for any language whatsoever. This has three interrelated reasons. First, there is no well-defined descriptive language which would allow the researcher to represent the meaning of some lexical unit, be it simple or compound; the most common practice is still to paraphrase it by an expression of the same language. Second, there is no reliable and easily applicable method of determining the lexical meaning of some unit; the most common way is to look at a number of occurrences in ongoing text and to try to understand what it means. Third, the relation between a particular form and a particular meaning is hardly ever straightforward; this is strikingly illustrated by a look at what even a medium-sized English dictionary has to say about the meaning of, for example, *on*, *sound*, *cast*, or (*to*) *put*. As a rule, there is not just one lexical meaning, but a whole array of uses, which are more or less related to each other. This is not merely a practical problem for the lexicographer; it also casts some doubt on the very notion of 'lexical unit' itself (*see* Lexical Semantics).

Making Dictionaries

Lexicographers often consider their work to be more of an art or a craft than a science (see Svensén, 2009). This does not preclude a solid scientific basis, but it reflects the fact that their concrete work depends largely on practical skills such as being 'a good definer,' on the one hand; and that it is to a great extent determined by practical, often commercial, concerns, on the other. Dictionaries are made for users, and they are intended to serve specific purposes. Their compilation requires a number of practical decisions.

Which Lexical Units Are Included?

Languages are neither well defined nor uniform entities; they change over time, and they vary with factors such as place, social class, or area talked about. A great deal of this variation is lexical. It is not possible nor would it be reasonable to cover this wealth in a single dictionary. Large dictionaries contain several hundred thousands of 'entries'; since idiomatic expressions are usually listed under one of their components (such as *to kick the bucket* under (*to*) *kick*), they contain many more lexical units, perhaps up to one million. But, even so, they are by no means exhaustive. The second edition of the *Deutsches Wörterbuch* (see Section Shortcomings), the largest dictionary of German, covers less than 25% of the lexical units found in the sources, and these sources are quite restricted themselves.

Which Lexical Properties Are Described?

Just as it is impossible to include all lexical units of a language in a dictionary, it is neither possible nor desirable to aim at a full description of those which are included. Since a dictionary is traditionally a printed book, the graphematic properties of the unit (its 'spelling') are automatically given. Among the other defining properties, meaning is traditionally considered to be most important. Samuel Johnson's dictionary from 1755 (see Section Shortcomings) defines 'dictionary' as "A book containing the words of any language in alphabetical order, with explanations of their meaning." But Johnson also noted which syllable carries the main stress, and he gave some grammatical hints. In general, however, information on phonological properties was rare up to the end of the nineteenth century, and information on grammatical properties is usually still very restricted in nonspecialized dictionaries. But there are, of course, also dictionaries which specifically address these properties as well as some of the nondefining properties associated with a lexical entry, such as its origin (etymological dictionary) or, above all, its equivalent in other languages ('bilingual dictionary').

What Is the Description Based Upon?

Usually, two types of sources are distinguished: 'primary sources' are samples of text in which the unit is used, 'secondary sources' refer to prior work of other lexicographers (and lexicologists). In fact, there is a third source, normally not mentioned in the theory of lexicography: This is the lexicographer's own knowledge of the language to be described, including his or her views on what is 'good' language. In practice, the bulk of a new dictionary is based on older dictionaries. This is always immoral and often illegal, if these are simply copied; but on the other hand, it would be stupid and arrogant to ignore the achievements of earlier lexicographers.

How Is the Information Presented?

A dictionary consists of lexical entries arranged in some conventional order. Normally, an entry combines several lexical units under a single 'headword'; thus, all lexical units which include the word *put* may be listed under this headword, forming a kind of nest with an often very complex microstructure. We are used to alphabetically ordered dictionaries; but there are other possibilities, for example, by thematic

groups or by first appearance in written documents. Languages without alphabetic writing require different principles; in Chinese, for example, entries are usually arranged by subcomponents of the entire character and by the number of strokes.

These four questions can be answered in very different ways, resulting in very different types of dictionaries (see the survey in Hausmann et al., 1991: 968–1573).

History

The first lexicographic documents are lists of Sumerian words (up to 1400) with their Akkadian equivalents, written in cuneiform script on clay tablets about 4700 years ago. The practice of compiling such word lists was continued throughout Antiquity and the Middle Ages; thus, the oldest document in German, the *Abrogans* (written around 765), is an inventory of some Latin words with explanations in German. Usually, these 'glossaries' did not aim at a full account of the lexicon; they simply brought together a number of words which, for one reason or another, were felt to be 'difficult,' and explained them either by a more familiar word in the same language or by a translation. Words were ordered alphabetically, by theme, or not at all. But there are also more systematic attempts, such as the *Catholicon*, a mixture of encyclopedia and dictionary which, compiled around 1250, was the first printed lexical work in Europe (Mainz 1460).

In the sixteenth century, two developments led to major changes. The first of these was the invention of printing by Gutenberg. By 1500, virtually all classical authors were available in print, thus offering a solid basis for systematic lexical accounts of Latin and Greek, such as Calepinus' *Dictionarium* (Dictionary) (1502), soon to be followed by two early masterpieces: Robert Etienne's *Dictionarium seu Latinae Linguae Thesaurus* (*Dictionary or Thesaurus of the Latin Language*) (Paris 1531) and Henri Etienne's *Thesaurus Graecae Linguae* (*Thesaurus of the Greek Language*) (Paris 1572). The second major development was the slow but steady rise of national languages. Since early Italian, French, English, or German were hardly codified, a major aim of the first dictionaries in these languages was to give them clear norms. In some countries, national Academies were founded to this end. The outcome were dictionaries with a strongly normative, often puristic, stance, such as the *Vocabulario degli Academici della Crusca* (*Vocabulary of the Members of the Academia della Crusca*) (Venice 1612), the *Dictionnaire de l'Académie Française* (*Dictionary of the French Academy*) (Paris 1694), and the *Diccionario de autoridades publicado por la Real Academia Espanola* (*Dictionary of the Authorities published by the Royal Spanish Academy*) (1726–1739). The bulk of lexicographic work, however, was always done by enterprising publishers and engaged individuals, such as Dr Samuel Johnson. Helped by six assistants, he produced *A Dictionary of the English Language* (London 1755), the first scholarly description of the English vocabulary, in less than 8 years. It surpassed all its predecessors, including Bailey's *Dictionarium Britannicum* (*British Dictionary*) from 1736, which Johnson took as his point of departure, by the systematic use of quotations, taken from the 'best writers,' and by his brilliant, sometimes somewhat extravagant, definitions (not everybody would dare to characterize patriotism as 'the last refuge of a scoundrel').

The rise of historical–comparative linguistics in the early nineteenth century led to an enormous increase in grammatical and lexical knowledge. The first dictionary which tried to cover this knowledge was the *Deutsches Wörterbuch* (*German Dictionary*) by Jacob Grimm and (to a lesser extent) his brother Wilhelm. Its first fascicle appeared in 1852, after about 10 years of preparatory work, in which the Grimms were helped by about 100 scholars providing excerpts ("covering my desk like snowflakes," Jacob Grimm). At that time, it was already clear that the original plan of six to seven volumes, to be finished within 10–12 years, was unrealistic. The Grimms finished only letters A to (most of) F, and the final folio volume (of altogether 32) appeared in 1961. This long duration, as well as the varying talents and preferences of the contributors, has led to many inconsistencies; some entries got out of balance (no less than 60 folio pages are devoted to the single word *Geist* (mind, spirit, ghost)); still, it is an incommensurable source of lexical information.

The work of the Grimms inspired a number of similar ventures, such as Emile Littré's masterly *Dictionnaire de la langue française* (*Dictionary of the French Language*) (1863–73), which is much shorter, but also much more consistent; Matthias de Vries and his numerous successors' voluminous *Woordenboek der Nederlandsche Taal* (*Dictionary of the Dutch Language*) (1864–1998), and finally *A New English Dictionary on a Historical Basis* (1884–1928), generally referred to as the *Oxford English Dictionary* (OED). It was initiated in 1857 by the philologist and churchman Richard Trench; in 1860, members of the Philological Society started to collect excerpts; in 1879, the Clarendon Press appointed James Murray as the Principal Editor. The first fascicle appeared in 1882, and the whole work was completed in 1928, 13 years after Murray's death. More than 200 scholars were involved in its production, more than 2000 people are known to have contributed excerpts. The OED is not without flaws, even in its revised edition, which appeared in 1989 in print and in 1992 on CD-ROM; but among all attempts to describe the lexicon of a language, it comes closest to falsify what Dr Johnson stated in the preface to his own dictionary: "Every other author may aspire to praise; the lexicographer can only hope to escape reproach." (For a comprehensive survey of lexicographic work across languages, see Hausmann et al., 1991: 1679–2710, 2949–3119.)

Shortcomings

At the end of the twentieth century, the art of making dictionaries had reached a remarkable degree of perfection. But two reservations are in order here. First, this perfection only holds for about 5% of the about 7000 languages of the world; for all others, we only have only more or less fragmentary and often unreliable accounts of their lexical repertoire – if any. Second, even the best dictionaries have individual as well as systematic shortcomings. Lexicographers are human beings, and thus, there are individual errors, poor definitions, or arguable decisions that make their achievements imperfect and popular targets of criticism. But there are

also deficiencies that result from the very nature of the printed book itself. These include:

1. Size restrictions. In a comprehensive corpus of present-day German, we find about five million lexical units (not counting dialects); the largest dictionary of present-day German describes about 170 000 lexical units – in 10 volumes, thus less than 5% of the words that are actually used.
2. Inadequacies of description. As a rule, the information given in a dictionary grossly underspecifies the various properties of a lexical item. Phonological properties are characterized by a phonetic script, such as the International Phonetic Alphabet (IPA); but from such a transcription, most users can only guess what a word really sounds like. The grammatical properties of a word are normally much more complex than what can be described by simple indications such as 'noun' or 'transitive verb.' Hardly anyone understands which familiar English words are targeted at by the meaning descriptions "Before the time in question, beforehand; by now; as early or as soon as this" or "diversion, amusement, sport; also, boisterous, jocularity or gaiety, drollery. Also, a source or cause of amusement or pleasure." (This is how the OED – the best dictionary of the world – describes the main readings of *already* and *fun*, respectively.)
3. Exclusion of overarching lexical properties. The vocabulary of a language is not just an aggregation of individual words, it is a highly complex structure, defined by the many relations which obtain between the phonological, grammatical, and semantical properties of all words (*see* Lexicon). A traditional dictionary with its focus on individual lexical units can hardly account for this aspect of a lexicon.

Since these deficiencies reflect inherent limits of a printed book – and this is what we normally understand a dictionary to be – it is practically impossible to overcome them within the limits of that format. Other formats must be developed to compile and to present lexical information. It is the advent of digital methods which begins to make this possible.

Digital Lexicography

Over the last decades, digital methods began to change, and in fact to overthrow, traditional lexicography. Five main lines in this development can be distinguished. First, it was necessary to create specific dictionaries for various tasks of natural language processing, such as machine translation, man–machine dialogue systems, or automatic text analysis. These 'dictionaries' are sometimes based on a machine-readable version of an existing printed dictionary. But the lexical information is stored on the computer, and it is given by some formal representation, rather than in a natural language. In the event, this led to completely new ways to describe the meaning of lexical units: the classical paraphrase in a few words was replaced by a netlike structure, in which various elements which play a role for that word are connected by semantic relations such as 'is a,' 'is a part of,' 'entails,' etc. The best known of these systems are WordNet (developed in Princeton) and FrameNet (developed in Berkeley).

Second, there is an increasing number of databases, mostly accessible via the Internet, which provide computer-generated information about various lexical properties of a language, for example, the frequency of words (or parts of words such as graphemes and morphemes) or of word co-occurrences. Good examples are the databases at the Brigham Young University (http://view.byu.edu) for English, the *Centre National de resources textuelles et lexicales* (National Center for textual and lexical resources) (www.cnrtl.fr) for French or *Wortschatz* (Vocabulary) (http://corpora.informatik.uni-leipzig.de), which provides statistical information about word frequencies and collocations in about 230 languages. Such databases often exploit huge corpora, sometimes in the range of billions of words. But these corpora are mostly not carefully compiled but based on material that happens to be available in digital form (like recent newspapers or Wikipedia texts), and, more importantly, their analysis is based on bare written word forms, without looking at the meaning. Thus, homographs such as *bear* and *bear* or *left* and *left* are considered as the same lexical item; this may limit their reliability for lexicology and lexicography.

Third, existing printed dictionaries were transferred to the computer and adapted to the purposes of the common user. Such a transfer offers several advantages: Search is faster and more exhaustive; the information is easier to revise and to update; and it is possible to add information not available in book format, for example, spoken sound instead of phonetic transcriptions. There are an increasing number of mostly free web sites which contain links to or even host dozens of – mostly bilingual – dictionaries, such as www.lexilogos.com, dictionary.cambridge.org, or www.dict.cc. They offer a cheap, fast, and easy source of lexical information, whose quality, though, varies considerably. Revisions and amendments, if there are any, are largely in the hands of volunteers.

Fourth, computers were increasingly used as tools in the *production* of a new dictionary. Rather than having a number of people read through books and newspapers and make excerpts of all occurrences which look interesting, it is now possible to compile huge text corpora that cover all varieties of a language, to scan these texts for all occurrences of words or word combinations, to sort these occurrences by various criteria, to link them to other occurrences, to add as much context as needed, etc. (Atkins and Rundell, 2008; *see* Corpus Linguistics). The first OED was based on about five million excerpts, mostly handwritten on paper slips. A computer can easily process corpora of several hundred million words, i.e., several hundred million occurrences; new sources can rapidly be added. This allows a much broader and much more representative coverage of a lexicon than ever. But electronic corpora only provide the raw material; it still awaits lexical analysis. This analysis can be facilitated by computer tools, also; but no computer can tell us what a word means in a particular context. If only 1 min is devoted to each occurrence in a one billion word corpus, it would take 100 lexicographers 100 years to go through it. This means that printed dictionaries can never reflect the wealth of information accessible in large corpora, because they are outdated when the lexicographers come to the end of their analysis. Therefore, the most recent version of the OED is no longer a printed book but a regularly updated and expanded computer

dictionary that is only searchable over the Internet (www.oed.com).

The fifth and last development is complex *digital lexical systems* which integrate various sources of information. They consist of (1) a computer-accessible and expandable corpus, (2) a set of tools, which allow, for example, not only the search for certain items but also statistical analysis or the determination of the first occurrence, or just look up of a number of occurrences that illustrate the use of an unknown word, and (3) a selective but steadily proceeding lexical analysis of the corpus by experienced lexicographers; in that respect, they are in line with the established lexicographical tradition. According to the need of the users and the means that are available, this analysis may vary in depth and coverage. Thus, it is possible to add spoken forms in various dialects, information about word classes, or the semantic analysis of some subset of lexical units, say all prepositions or all morphologically simple verbs. Similarly, translation equivalents can be added. Unlike printed dictionaries, such a lexical retrieval system, such as the *Digitale Wörterbuch der Deutschen Sprache* (*Digital Dictionary of the German Language*) (www.dwds.de), will never come to an end, it is steady work in progress to which many can contribute and which will give us a deeper and broader understanding of the lexicon than any other method.

See also: Corpus Linguistics; Encyclopedias, Handbooks, and Dictionaries; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Lexicon.

Bibliography

Atkins, Rundell, 2008. The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford.

Cruse, D. Alan, Hundsnurscher, Franz, Job, Michael, Lutzeier, P.R. (Eds.), 2005. Lexikology – Lexicology. de Gruyter, Berlin.

Hausmann, Franz Joseph, Reichmann, Oskar, Wiegand, Herbert Ernst, Zgusta, Ladislav (Eds.), 1991. Wörterbücher – Dictionaries – Dictionnaires. De Gruyter, Berlin.

Jackson, Howard, 2013. The Bloomsbury Companion to Lexicography. Bloomsbury, London.

McArthur, Tom, 1986. Worlds of Reference. Lexicography, Learning and Language from the Clay Tablet to the Computer. Cambridge University Press, Cambridge, UK.

Svensén, Bo, 2009. A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge University Press, Cambridge.