Chapter 4 in: Sylviane Granger and Magali Paquot (eds., 2012) *Electronic Lexicography*. Oxford University Press.

Corpus Evidence and Electronic Lexicography

Patrick Hanks
Research Institute for Information and Language Processing, University of Wolverhampton;
University of the West of England, Bristol

Abstract

Corpus evidence opens up radical possibilities for lexicography. Traditional dictionaries tend to distort meaning because they pay insufficient attention to phraseology. Corpus evidence provides massive evidence for both normal and abnormal phraseology. For the proper analysis of meaning in text, a new theory of language is needed, along with new kinds of dictionaries. These will take account of prototype theory and stereotype theory and apply them to both phraseology and word meaning. Linguistic phenomena such as these are probabilistic, not deterministic. Electronic lexicography in future will have to take account of phraseological norms and statistical analysis of word use. Conventions of word meaning need to be associated with the word in its normal contexts (i.e. conventions of usage) rather than merely with the word in isolation. The chapter concludes by considering the pros and cons of Wiktionary as a possible model for electronic lexicography of the future.

Keywords: corpus evidence, historical dictionaries, bilingual dictionaries, collocations, statistically significant co-occurrences, phraseological prototypes, syntagmatic preferences, idiomatic phraseology, salient probabilities, framenet, Corpus Pattern Analysis, Wiktionary

1. Introduction

This chapter starts by examining the impact that corpus evidence, i.e. electronic evidence of the way words are actually used, has had, is having, and will most probably continue to have on different traditions of lexicography. Electronic lexicography opens up all sorts of radical possibilities that were closed to traditional lexicography: new kinds of evidence, new modes of description, new ways of organizing evidence, new possibilities for exploiting database structure and hypertext links, and the need for new theoretical foundations. Most important among these new possibilities, from a practical lexicographical point of view, is the opportunity to build hypertext databases showing explicit links between word senses and patterns of word use. Allied to this are opportunities to study the situations or 'frames' in which words are actually used. These and other opportunities offered by electronic lexicography raise interesting issues at both a practical and a theoretical level.

The space constraints of the printed book have been removed by the Internet—but so has the commercial incentive to manufacture dictionaries as physical objects for sale to users. At present, lexicography is in transition: publishers, nervous about future commercial prospects, are wary of investing in large-scale innovations, just at a time when such innovations are most needed. At the same time, funding agencies and their advisers have not yet been convinced that major innovations such as Frame Net and Corpus Pattern Analysis would justify the large-scale research investments that would be needed to bring them to completion and yield practical benefits.

In this chapter at least, optimism must prevail. Let us assume that somehow, somewhen, the need for major lexicographical initiatives will be recognized and will be funded—even though, in the age of the Internet, the deliverables are unlikely to be in the form of traditional printed books. So the chapter continues by asking what needs to be done and concludes by suggesting a context in which it might be done.

2. Three traditions of lexicography

Scholarly dictionaries on historical principles. Traditionally, the task of lexicography was seen, at least among scholars and literati, as being to compile an inventory of all the words in a given language and to trace their origins and their semantic development. Samuel Johnson (1755), in his famous definition of a lexicographer as "a harmless drudge", saw the task of the lexicographer as being not only "detailing the signification of words", but also "tracing the original [of words]" [my emphasis]. From Johnson's day up until very recently, it was assumed that origins determine meaning. That is to say, if a word changes its meaning over time, it was assumed that the older meaning was somehow more correct than any more recent development. This was an underlying assumption for three hundred years, and it influenced (among others) the Philological Society during their deliberations in the 1850s, which led to the foundation of "a New English Dictionary on Historical Principles" [my emphasis again], later to become better known as the *Oxford English Dictionary* (OED). The currently ongoing magnificent blend of scholarship and technology that is creating the third edition of OED (OED3), seeks to elaborate rather than challenge the basic historical principles ad philosophical assumptions of the Philological Society and in particular of James Murray, first editor of the first edition of OED, during the 19th century. OED3 is an exemplary on-line historical lexicographical product, with the results of lexical research being made available more or less immediately, rather than years later on completion of the whole book. It is no longer constrained by the requirement to work in alphabetical order. However, quite rightly, it does not seek to make radical alterations to the received 19th-century model of scholarly lexicography.

Dictionaries on historical principles have an important integrative social role to play in a culture, *sub specie aeternitatis*, in explaining the changes of meaning that so many words in all modern languages have undergone. Roots are culturally important and scholars and literati need to know about them. However, knowing about roots is not the same as knowing about meaning. Dictionaries in a single volume aimed at practical users among the general public are a different matter. Is it truly helpful to a general user to be told that sense 1 of ***camera*** is "a small vaulted room", and in particular "the treasury of the papal curia"? It is a matter for astonishment that throughout the 20th century many one-volume monolingual English dictionaries arranged their senses in historical order, and this practice persists to this day in America's favorite dictionary (the Merriam-Webster Collegiate series). It would seem that, unless you already know what a word means in contemporary English, you cannot use such a dictionary with confidence to discover its meaning.

Recognition that people might want to use a dictionary to find out what words mean, rather than where they came from, was slow to establish itself in the face of benighted reverence for erudition. Funk and Wagnall's *Standard Dictionary of the English Language* (1894-97) was the first large-scale dictionary to set out deliberately to reject historical principles and instead record the current meaning of words, followed eventually by the *American College Dictionary* (1947), the *American Heritage Dictionary* (1968), and, in Britain, by the *Hamlyn Encyclopedic World Dictionary* (1971), and *Collins English Dictionary* (1979). However, the lexicographers working on these dictionaries, attempting to follow synchronic semantic principles, ran into a difficulty. In the absence of large bodies of evidence, how were they to identify the most common meaning of a polysemous word? This problem is particularly acute for light verbs such as *take* and *bear*, but it affects many less common words as well, such as *launch*, *spoil,* and *dope* (to select at random just three of many hundreds of examples that could be given). And when very large bodies of evidence did eventually become available, for example the *British National Corpus* and the *Bank of English*, there were some surprises in store. Our intuitions as native speakers about normal and most frequent senses of words turned out to be utterly unreliable. Social salience (i.e. the most frequent sense or use of a word) and cognitive salience (its most literal and obvious meaning) are independent variables (see Hanks 1990, 2010).

Two other, equally important traditions, which must be mentioned here, are bilingual dictionaries and dictionaries for foreign learners.

Bilingual dictionaries. At least since Colin Smith's *Collins Spanish-English Dictionary* of 1971 and the *Collins-Robert French-English Dictionary* by Beryl T. (Sue) Atkins and Alain Duval (1978), bilingual lexicographers have attempted to give practical implementation to the long-recognized fact that aiming at literal word-for-word translation between languages is a naive goal that leads to errors—often, ludicrous errors. Such lexicographers have paid increasing attention to phraseology. Their purpose in doing this was not to explore the relationship between word meaning and word use at a theoretical level, but rather to discover commonplace idiomatic expressions that cannot be translated literally, word for word, from a source language into a target language. To do this, they compiled large frameworks consisting of typical phraseology for each word in each language. The lexicographers would then work in pairs (each pair consisting of a native speaker of each language) to establish idiomatic and pragmatic equivalents in the target language for the phrases in the source-language framework. In this heroic endeavour, they were hampered by lack of evidence. What exactly are the typical phrases associated with each word in a particular language? Corpus evidence, as we shall see, was to provide at least a partial solution to this problem.

Dictionaries for foreign learners: A third, equally important lexicographic tradition began in 1942 with the publication in Japan of A. S. Hornby's pioneering *Idiomatic and Syntactic English Dictionary* (ISED, 1942, re-published in 1948 by Oxford University Press as the *Advanced Learner's Dictionary*). It was not until 1978 that a serious rival to this wonderful dictionary (by now re-named *The Oxford Advanced Learners' Dictionary of Current English*) appeared: this was the *Longman Dictionary of Contemporary English* (LDOCE). Hornby's original aim was to create a work that would help learners to use the syntactic patterns and idiomatic phraseology of English with reasonable accuracy when writing and speaking. In other words, ISED was intended as an aim for language production and would ignore words needed only for 'decoding' purposes (i.e. reading skills). Hornby's verb patterns have been superseded now, but in their day they provided revolutionary insight into verb valency in English. Until the corpus revolution of the 1980s and 90s, they stood the test of time remarkably well, and they must have helped literally millions of learners of English world-wide during the mid twentieth century. The evidence upon which they were based consisted largely of introspection by Hornby and his colleagues.

In the second and subsequent editions of OALDCE, this purely productive aim was watered down. Hornby and his successors found it impossible to resist the criticism that a dictionary ought to explain words that learners would be likely to encounter during their reading and listening activities in English, as well as words that they would actually use. Thus, many thousands of words were added to the second and subsequent editions—words presenting no serious difficulty in their idiomatic and syntagmatic behaviour: words that learners might be unfamiliar with (and for which they would therefore turn to a dictionary for an explanation) but would be unlikely ever to use in their spoken and written use of English. It was this dual aim—decoding as well as encoding—with which LDOCE and subsequent English learners' dictionaries set themselves the task of competing. Later dictionaries such as Cobuild, CIDE (*Cambridge International Dictionary of English*; subsequently retitled CALD (*Cambridge Advanced Learners Dictionary*)), and MEDAL (*Macmillan English Dictionary for Advanced Learners)* were to enlist the assistance of corpus data in pursuance of this aim.

3. The impact of corpora on lexicography

The impact of corpus evidence on lexicography is described more fully in Hanks (2009). Early electronic corpora (Brown, LOB) had little impact on lexicography, despite being consulted by major dictionaries (in particular, the *American Heritage Dictionary* (first edition 1968) and the *Longman Dictionary of Contemporary English* (LDOCE; first edition 1978). The reason was simple: these were each corpora of only one million words—corpora so small that it was impossible to distinguish statistically significant co-occurrences of words from chance co-occurrences. This implies that in order to discover and organize word meanings, it is necessary to study textual evidence and in particular collocations. This hypothesis is one that was first proposed by J. R. Firth (1957a,b) and was developed by the late John Sinclair from his earliest, prophetic published work (Sinclair, 1968) to his posthumously published essay entitled 'Defining the Definiendum' (Sinclair 2010). His life's work was largely devoted to the development of a theory of language and meaning based on the empirical investigation of the collocational preferences of words. In Sinclair (1987) he asked:

> How common are the phrasal verbs with *set*? *Set* is particularly rich in making combinations with words like *about, in, up, out, on, off,* and these words are themselves very common. How likely is *set off* to occur? Both are frequent words; [*set* occurs approximately 250 times in a million words and] *off* occurs approximately 556 times in a million words.... The question we are asking can be roughly rephrased as follows: how likely is *off* to occur immediately after *set*? ... This is 0.00025 × 0.00055, which gives us the tiny figure of 0.0000001375 ... The assumption behind this calculation is that the words are distributed at random in a text. It is obvious to a linguist that this is not so, and a rough measure of how much *set* and *off* attract each other is to compare the probability with what actually happens... *Set off* occurs nearly 70 times in the 7.3 million word corpus. That is enough to show its main patterning and it suggests that in currently-held corpora there will be found sufficient evidence for the description of a substantial collection of phrases....

The first fruits of the Sinclairian approach to corpus-driven lexicography emerged with the first edition of the Cobuild dictionary (1987), based on an initial corpus of approximately 7.3 words, which by the time of publication had grown to 18 million words—just large enough for the main patterns of collocation associated with each word to be perceived through what J.R. Firth had called "the mush of general goings-on":

> "We must separate from the mush of general goings-on those features of repeated events which appear to be part of a patterned process." —J. R. Firth (1950)

A special issue of the *International Journal of Lexicography* (21:3, September 2008) was devoted to the intellectual legacy of John Sinclair from a variety of lexical viewpoints.

Thus, the first major impact of corpora on lexicography was on dictionaries for foreign learners. Subsequent newly compiled learners' dictionaries (CIDE, Macmillan) were also corpus-based, though none were corpus-driven in the way that Cobuild was. In due course complete recensions of the leading English dictionaries for foreign learners, OALD and LDOCE, were prepared on the basis of corpus evidence, though for marketing reasons the distinction between a dictionary as an encoding aid and as a decoding aid came to be fudged by the publishers and hence by the lexicographers. Comparing entries in pre-corpus editions of OALD and LDOCE with entries for the same words in post-corpus editions of the same dictionaries and in Cobuild and other corpus-driven dictionaries reveals that corpus-based dictionaries—even dictionaries based on different corpora—have tended to converge in what they say about the language, compared with pre-corpus dictionaries, as described for example in Atkins and Levin (1991).

Subsequently, the one-volume *New Oxford Dictionary of English* (1998; subsequently rechristened the *Oxford Dictionary of English, ODE,* 2001) made extensive use of corpus evidence to compile a brand-new account of contemporary English for use by native speakers. To date, *ODE* is the only monolingual dictionary of English for native speakers to be corpus-based. The *New Oxford American Dictionary* is an Americanization of it. In other languages, the situation is rather different: for example, major corpus-based dictionaries of languages as different as Danish, Modern Greek, and Malay have been published.

The impact of corpus data on synchronic lexicography since 1987 (the date of publication of Cobuild) has been overwhelming. At last lexicographers have sufficient evidence to make the generalizations that they need to make with reasonable confidence. We can now see that pre-corpus lexicography was little more than a series of stabs in the dark, often driven by historical rather than synchronic motives. In word after word, pre-corpus lexicographers (consulting their intuitions and a bundle of more or less unusual citations collected by citation readers) failed to achieve the right level of generalization regarding the conventions of present-day word meaning in a language, as can be seen by attempting to map the old definitions onto the new evidence. Of all the many possible uses and meanings that a word might have, lexicographers now have better chances of selecting the ones that are actually used and of writing reasonably accurate descriptive definitions of commonly used words.

Large corpora provide monolingual lexicographers with sufficient evidence to decide what to include and (more importantly) what to leave out. Corpus evidence contributes to the never-ending task of improving the accuracy of explanations and provides evidence for the pragmatic uses of words and phrases, which had been largely neglected in traditional dictionaries. Large corpora provide evidence for 'local grammar' or 'valency'—the syntagmatic structures in which each word is normally used (as opposed to speculations about how it might possibly be used). Above all, they provide evidence for collocations—the preferences that words have for the company of certain other words. This is a subject that could not be studied empirically before the evidence of large corpora became available, together with statistical techniques for the analysis of word associations (see Church and Hanks 1989; Kilgarriff 2005).

Turning back now to bilingual dictionaries, a careful comparison of the *Collins-Robert French-English Dictionary* with the later *Oxford-Hachette French Dictionary* (Corréard and Grundy, 1994), in which Atkins played a major role as an adviser, will give a slight indication of how corpus evidence is able to refine and extend the 'framework' approach to bilingual lexicography. It should be borne in mind that in the early 1990s, corpus evidence was extremely scarce. Nevertheless, the Oxford-Hachette shows a more focused (but not larger) selection of phraseology than its Collins-Robert predecessor. An entry for a commonplace English word such as *day* includes not only an indication of the French semantic distinction, not made in English, between *jour* and *journée*, but also over eighty model phraseological equivalents, including:

> the day before = *la veille*
> to come on the wrong day = *se tromper de jour*
> before the day was out = *avant la fin de la journée*
> at close of day = *à la tombée du jour* [notice that in French the day falls, whereas in English it is nights that fall]
> it was a hot day = *il faisait chaud*
> in his/her younger days = *dans sa jeunesse*
> in those days = *à cette époque*

Notice that in the first and the last three examples given here, there is no French word at all that can be literally translated as 'day'. The meaning is conveyed by other means. The earlier *Collins-Robert French Dictionary* has a larger but less well-selected collection of phraseology. What distinguishes the very different selection of phraseology in the *Oxford-Hachette French Dictionary* is improved selectivity—that is, corpus evidence has enabled the lexicographers to select phraseology that is more frequently used and therefore more likely to be useful to users. The lexicographical emphasis has shifted from the hopeless aim of covering all phraseological possibilities to the more realistic (and empirically well-founded) one of covering the most salient probabilities.

In the context of scholarly lexicography on historical principles, the impact of corpus evidence has so far been less dramatic. Such an impact may be expected when large historical corpora of language become available, enabling lexicographers to distinguish phraseology that seems unusual to us today because the norms of the language have changed from phraseology that was idiosyncratic to a particular writer, text, or small group of texts.

4. How will electronic monolingual lexicography of the future be different from traditional models?

Analysing word meaning in context is a skill that is still in its infancy. Many roads will be tried. In current dictionaries, other than (up to a point) Cobuild, context has been very largely neglected or at best regarded as a sort of optional extra. In dictionaries of the future, contextualization and phraseology will come to take centre stage. These dictionaries will be electronic products with hypertext structures and links, not printed books, nor the 'horseless carriages' that now pass for electronic dictionaries.

An underlying theme running through what has been said so far in this chapter is that word meaning can only be described accurately if the word is put into context. Part of the lexicographical task required of a lexicographer using a large corpus is to select contexts that are maximally general (and therefore offer maximum predictive power about the word's meaning in future, unseen contexts), while at the same time preserving a sharp semantic focus. "Context" here can have two, interrelated meanings: context of utterance (the real-world situation in which a word is uttered) and textual context ('co-text'; syntagmatic preferences for use within a certain syntactic structure along with the company of other words, i.e. significant valency and collocations).

Two kinds of context must be taken into account before the meaning of a word can be accurately described: context of utterance and syntagmatic linguistic context. Let us look at two theoretical and practical approaches that focus on each of these.

4.1 Semantic analysis based on context of utterance – Frame Semantics and FrameNet

Charles Fillmore has made at least three important contributions to linguistic theory with a semantic component: case grammar, frame semantics, and construction grammar, each of which represents a plank in a possible bridge between syntax and lexical semantics. He has always evinced an interest in meaning as well as syntax, and he was one of the first linguists to recognize the importance of prototype theory (see, for example, Fillmore 1975). His interest in semantics is associated with analysis of the lexicon, and for many years, during the development of the FrameNet project, he worked closely with the lexicographer Sue Atkins. He is one of the few American linguists to show awareness of European schools of linguistics. His published works regularly cite major European theorists such as Tesnière, Maurice Gross, Trier, and Helbig, among others, as well as contemporary American linguists.

In the following few paragraphs, I shall focus on Fillmore's frame semantics, from its source in case grammar to its practical realization in FrameNet. Frame semantics originated in case grammar (Fillmore 1968), in which every verb is identified as selecting a certain number of basic cases, which form its '**case frame**'. For example:

> **give** selects three cases: *Agent* (the person doing the giving), *Benefit* (the thing given), and *Beneficiary* (the person or entity that receives the Object);
>
> **go** selects two cases: *Agent* and *Path* (more specifically, subdivided into *Source, Path, Goal*);
>
> **break** selects three cases: *Agent, Patient* (the thing that gets broken), and *Instrument* (the object used to do the breaking, for example a hammer).

These cases may appear in different syntactic positions. Levin's examples (1, 2 below) show that the 'Patient' may appear both as the direct object of a causative verb and as the subject of the same verb used inchoatively.

1. *Janet broke the cup.*

2. *The cup broke.*
   —Examples from Levin (1993).

In Frame Semantics, frames are conceptual structures involving a potentially large number of lexical items, not just individual meanings of individual words. Fillmore (1982) says that Frame Semantics "offers a particular way of looking at word meanings", but then immediately goes on to say:

> By the term 'frame', I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits.

Thus, the claim is that to understand the meaning of a word, you need access to all the essential knowledge that relates to it. For example, to understand ***sell,*** you need to know about the 'frame' of commercial transactions, with *Seller, Buyer, Goods* [alternatively, *Service*], and *Money.* You also have to know about relations between *Money* and *Goods*; between *Seller, Goods*, and *Money*; between *Buyer, Goods* and *Money*; and so on.

> A word's meaning can be understood only with reference to a structured background of experience, beliefs, or practices, constituting a kind of conceptual prerequisite for understanding the meaning. Speakers can be said to know the meaning of a word only by first understanding the background frames that motivate the concept that the word encodes. Within such an approach, words or word senses are not related to each other directly, word to word, but only by way of their links to common background frames and indications of the manner in which their meanings highlight particular elements of such frames. —Fillmore and Atkins (1992)

This is rather different from the goal of understanding how a word is normally and idiomatically used in a language to make meanings, as we shall see in the next section.

FrameNet, the practical implementation of the theory of Frame Semantics, is work in progress. An interesting question is whether, in principle, it can ever be fulfilled. The answer is probably no, since there does not seem to be any very good reason to believe that the number of possible frames is finite.

Each **Frame** is populated by several **Lexical Units** and is supported by annotated corpus lines. A lexical unit is a pairing of a word with a meaning. **Frame Elements** are entities that participate in the frame. Different senses of polysemous words belong to different frames. A group of lexical units (words and multiword expressions—MWEs) is chosen as representative of a particular frame. For each lexical unit, a concordance is created from a corpus, and sample concordance lines are selected and annotated. Labels (names) are created for each of the Frame Elements. Fillmore (2006) discusses the example of the 'Revenge frame'. The following lexical items are identified as participating in this frame:

> verbs: *avenge, revenge, retaliate; get even, get back at; take revenge, exact retribution*

nouns: *vengeance, revenge, retaliation, retribution*

adjectives: *retaliatory, retributive, vindictive*

The Frame Elements are:

Offender, Injured Party, Avenger [may or may not be identical to the Injured Party], *Injury* [the offence], *Punishment*

The relationships are summarized as follows:

O has done I to IP; A (who may be identical to IP), in response to I, undertakes to harm O by P.

Despite the many examples in FrameNet taken from BNC, with extensive tagging of the thematic roles of lexical items, it would be a mistake to imagine that FrameNet is corpus-driven. The frame, its elements, and its lexical units are dreamed up by introspection; examples are imported after the event. No attempt is made to analyse systematically the meanings or uses of any given lexical item. FrameNet proceeds frame by frame, not word by word. As a result, there are many gaps, which will remain unfilled unless some member of the FrameNet team dreams up a relevant frame. Two examples will suffice, out of literally hundreds that could be mentioned.

- Most uses of the verb ***spoil*** denote destroying the pleasure of a special event, such as an outing or a party. Another large group of uses denote habitual pampering of a child. However, the only frame in FrameNet for this word is the rather rare Rotting frame (e.g. *I've got a piece of ham that'll spoil if we don't eat it tonight*), accounting for only about 3% of uses in BNC.

- There are two main uses of the verb ***admit***: one of them denotes saying something reluctantly; the other involves a pattern in which someone is taken into a hospital or a residential home for treatment or for care. At present only the first of these is covered by a FrameNet frame.

FrameNet is work in progress, so maybe, if FrameNet goes on long enough and if someone dreams up an appropriate frame, gaps like these will be plugged eventually. However, it seems unlikely that all such gaps will be plugged, for FrameNet does not have a target inventory of frames to create, does not have any criteria for distinction between frames, and does not have criteria for completion of the whole task. In other words, it is not based on systematic analysis of a target lexicon. Despite these somewhat negative comments, it must be emphasized that FrameNet is full of profound lexical and semantic insights that will repay careful study by anyone interesting in meaning in language.

A special issue of the *International Journal of Lexicography* (16:3, September 2003) was devoted to FrameNet and Frame Semantics.

Semantic analysis based on textual context - CPA

It is not the case that any bundle of authentic citations is acceptable by way of lexicographic evidence. Authenticity alone is not enough. In addition to authenticity, citations must be chosen to illustrate usage that is normal and idiomatic. Any corpus contains a small percentage of deliberate, authentic exploitations of normal usage. Lexicographers must be trained to recognize such exploitations for what they are, and not treat them as a reason for broadening the scope of definitions to the point where the focus is lost.

Lexicographers of the future will develop different approaches to different words, depending on the function of each word in the language. Some words, especially nouns denoting concrete objects, will need definitions of a fairly traditional kind that show how the word has concrete reference to a set of objects in the world. Many other words, on the other hand, including most verbs and adjectives and many nouns, especially abstract nouns, need to be explained in the context of their normal phraseology. The set of phraseological norms for any word may be regarded as a linguistic gestalt, which is associated with a complex of meanings and beliefs and may be used normally or exploited creatively, according to the speaker's or writer's needs. We are now brought face to face with a problem that has confronted lexicographers from time immemorial, or at least since dictionaries began. This is that the range of phraseology and meanings of certain words is of almost incredible complexity. How can any ordinary human language users carry lexical items of such enormous complexity in his or her head? I will hazard an answer to this question in a moment, but let me first give an example, which, I am sorry to say, will be rather space consuming. A corpus-based study of the verb *throw* yields the following observations, at a level of generalization suitable to account for the normal patterns of use of this verb found in BNC (which we may assume is a balanced and representative sample of English):

> *People throw hard physical objects like stones, bricks, and bottles* ***at*** *other people and things, typically but not necessarily with the intention of causing damage* [the preposition 'at' in this context intensifies the notion of intention to cause damage] – *people throw tomatoes and eggs at politicians to express contempt for them – terrorists throw bombs – soldiers throw grenades at the enemy – ball players throw balls* ***to*** *each other* [the preposition 'to' in this context intensifies the notion of cooperative behaviour; there is a whole complex of domain-specific secondary norms here] – *you can throw your hands or arms in the air (but they remain attached to your body) – you can throw your hat in the air (and you may fail to catch it as it comes down) – suicidal people throw themselves under trains, out of windows, into rivers or ponds – committed people throw themselves into an activity – you throw away (or throw out) things that you no longer want – if you are on a boat, you throw unwanted things overboard – if a proposition or argument (e.g. in a lawsuit) is unconvincing, the whole lawsuit or proposal may be thrown out by the judge or decision*

> *maker – a person may be thrown out of a place where they are not wanted – 'throwing out the baby with the bathwater' implies accidentally rejecting something of central importance at the same time as rejecting unwanted things associated with it.*
>
> *A person may throw* ***off*** *things like clothes and blankets – you can also throw off abstract things like moral restraints – a moving object may throw a person or object off – a person trying to find out something may be thrown off the scent – a person may be thrown off course, off balance, or off the scent – you can throw* ***in*** *your lot with someone else – you can throw something extra in (for good measure) with a set of things – a person may be thrown in at the deep end, in a new job for example – a person or physical object may literally be thrown* ***into*** *the air, for example by the force of an explosion – a person may be thrown into jail – a situation may be thrown into chaos, confusion, or turmoil – an idea may be thrown into doubt or into question – a concept can be thrown into relief by some contrasting event or concept – a defeated person throws in the towel – a troublemaker throws a monkey wrench (British: spanner) in the works.*
>
> Less frequent but nevertheless conventional phrases, with highly idiomatic meanings are the following (with cognitively salient collocates in bold): *bullies throw their* ***weight around*** *– powerful people throw their* ***weight behind*** *politicians or proposals – an event may throw* ***light on*** *a mystery – bad things throw* ***a shadow over*** *good things – evidence may throw* ***doubt on*** *a belief or hypothesis – boxers throw* ***punches*** *in a boxing match, while an aggressive person may throw* ***a punch*** *and start a fight – a group of people may throw* ***a party*** *– you might throw* ***down the gauntlet*** *or throw* ***out*** *a* ***challenge*** *to a rival – gamblers throw* ***dice*** *– an unstable or excitable person may throw a* ***tantrum****, a* ***fit****, or a* ***wobbly*** *– a reckless person throws* ***caution to the wind*** *– a situation may throw* ***up*** *new concepts or abstract entities – a person who has drunk too much or who has eaten something bad (or one who is exposed to extreme emotional distress) may throw* ***up****.*

Notice that, in this phraseological account of the verb *throw*, there are almost no explicit meaning statements—and yet the meaning of most of the phraseological prototypes listed here is probably clear enough to most people.

The answer to the question, 'How can we store all this in our heads?' may be that we don't. Probably, each member of a speech community has in his or her head only a subset of this whole gestalt (for purposes of vaguely recognizing meaning) and an even smaller subset for active use. It is a mistake to suppose that native speakers know the whole of their native language. They don't. Nobody can. A persons knows at best only a big chunk of their native language—the part that they have internalized and use for communicative purposes. It is perfectly possible to go through life without ever encountering, let alone using, a rare phrase such as 'throw down the gauntlet.' Your subset of the phraseology and meanings of *throw* and all other words is probably rather different from mine and those of other people that you know (not to mention the many millions of other Englsih speakers), but nevertheless there must be enough overlap for communication among us to be possible.

On the other hand, the enormous capacity of the human brain for storing experiences and the words associated with these experiences must not be underestimated. Maybe the various phrases embodying the verb *throw* that we use are stored separately in different places in our brains in the context of different communicative needs and associations, rather than as a homogeneous whole.

If this is right, at least one thing is wrong with this presentation here, namely that all the major syntagmatic components of this linguistic gestalt—*lexical gestalt* might be a better term—are listed *en masse*, in a quasi-rational sequence. This gives a misleading implication that all aspects of a lexical gestalt are or can be psychologically active at the same time in the mind of any one language user, or readily recalled to the conscious mind for purposes of exemplification and discussion. This is incorrect. In reality, the gestalts for such complex words are buried (in ways that we do not yet fully understand) deep in the subconscious. Different components are activated according to need and context of utterance. In particular, the text of any discourse—document or conversation—that leads up to the choice of the word ***throw*** sets preconditions such that only a tiny subset, consisting of particular, relevant aspects of the gestalt, are activated in any context. It is highly implausible that (as some psycholinguists have argued) in writing, reading, or conversation all the possible norms for a word are activated first and then the relevant one is selected by a speaker, writer, listener, or reader. A corpus may show us a fairly full set of what may be stored mentally, but it does not tell us how it is stored.

Whatever the psycholinguistic reality may be, the duty of the lexicographer is generally seen as being to report all the conventions of a language, not just some of them. Having said that, we must also acknowledge the impossibility of reporting all the conventions of a living language, which is constantly changing and developing and which may have many subdomains. The best we can hope for is to report all the common conventions of meaning and use, and to discover the general principles that relate one meaningful phrase to another and that govern the way in which conventional phraseology ad meaning may be exploited.

Different aspects of this complex gestalt are open to exploitation in various ways. A few examples of uses of ***throw*** may now be given to illustrate how phraseology is exploited and the kind of exploitation rules that are needed. The implicatures range from semi-literal to highly idiomatic or metaphorical.

Throwing a physical object *at* something, as we have denotes a volitional human action with the intention (not necessarily successful) of causing harm or damage. This notion is exploited in 1, a metaphor where the brick in question is not a physical object at all, though the intention to cause damage is clearly present.

1. Worldwide, the economy has continued to come on stronger than almost anyone forecast, which is why European central bankers agreed to **throw** another brick **at** it yesterday. —(BNC) *Independent*, electronic edition of 1989-10-06: Business section.

Why would central bankers want to damage the world economy? Because, according to writer of 1, in October 1989 the world economy was 'running too fast' and 'liable to overheat' (two fairly conventional metaphors used in the domain of economics).

Throwing bricks, throwing stones, and *throwing punches* are expressions that have approximately equal salience in the BNC. However, *throwing punches* is more often used metaphorically (and may be regarded as always somewhat metaphorical, for reasons explained in the next paragraph).

2. Punches **were thrown** outside the Queen's Head Hotel in Bishop Auckland —(BNC) *Northern Echo* [Date not given].

In order to understanding the meaning in 2, the reader needs to take account of the semantic types of the collocates. A stone is a physical object; throwing a stone is a physical event. A punch, however, is itself an event, not a physical object. In 4, therefore, the verb ***throw*** is semantically light: what is thrown is not a physical object but an event. The meaning is that various people literally and physically punched each other, not that some physical object was impelled through the air. However, this light-verb use may itself be exploited metaphorically, as in 3.

3. With the mass media now part of everyday life and with arguments about bias and balance commonplace, the modern British subject is not likely to succumb to some **Saddam sucker punch thrown** by the third party from the corner of the living room. —(BNC) *Marxism Today*, [Date not given].

The reference in 3 is to a possible aggressive remark, rather than to physical violence. In boxing, a ***sucker punch*** is a deceptive punch thrown in a way that deceives an inexperienced fighter, but here the domain is politics, not boxing. Instead, the expression denotes a deceptive and aggressive, potentially destructive remark. As readers, we deduce this interpretation from other collocates in the context: "arguments about [something]" and "the corner of the living room" are not compatible with a literal fight or boxing match, but they are compatible with aggressive remarks. The throw-away allusion to Saddam [Hussein] reinforces the notion of a deceptive remark. So this semantically dense sentence is highly metaphorical. The collocates and their semantic types determine the interpretation of the target word, *throw*.

To take another example, *throwing a shadow* likewise has a cline of metaphoricity, from light-verb uses to highly metaphorical references. In 4, the street light is a physical object, and shadows are visible objects (even though they have no physical substance). This is, then, an almost literal expression; the only thing about it is that is idiomatic is the choice of throw as a light verb to denote what is in effect a visual perception. In 5, on the other hand, the collocates *hindsight* and *retrospective* (among other collocates in this fragment) invite a metaphorical interpretation of the ***shadow*** that is thrown over past people and events.

4. The street light **threw strange shadows** among the hoardings. —(BNC) W. B. Herbert, 1992. *Railway ghosts and phantoms*.

5. There are dangers involved in the writing of contemporary history quite apart from the standard objection that distance and hard evidence are required if a true perspective is to be gained. Hindsight **throws a retrospective shadow** over people and events which distort light and shade as they were actually perceived at the time. The period of the Attlee governments of 1945-51 was particularly prone to retrospective retouching by the ideologically driven. —(BNC) Peter Hennessy, 1990. *Cabinet*.

Let us now turn to some graphic metaphorical uses of some of the phrasal verbs formed with ***throw***, starting with ***throw something overboard***. This nautical expression is stronger and more expressive even than *throw something away* and *throw something out*. If you throw something away or out, in the short term you still have the option of going to the bin and retrieving it. But if someone on a ship throws something overboard, it is lost irrevocably, for ever. This fact, coupled with the salience of nautical expressions in general in figurative English—a by-product of the important role that the sea has played in English history and in the spread of the English language—means that it is not surprising that this expression is often used metaphorically. 15 out of the 35 uses in BNC of the expression ***throw [something] overboard*** are metaphorical.

6. Emanuel Shinwell, who has never changed his mind on this issue, was clear in 1918 about the wrong-headedness of destroying the people's grammar schools while leaving unscathed the privileged Public Schools: 'We were afraid to tackle the public schools to which the wealthy people send their sons, but at the same time are ready to **throw overboard** the grammar schools which are for many working-class boys the stepping-stone to the universities and a useful career.' —(BNC) Harry Judge, 1984. *A Generation of Schooling*.

6 is a conventional metaphor. Less conventional is the metaphor in 7, with another phrasal verb, which has a somewhat similar meaning, namely *throw out.* The reader may feel that 'throw out' is a somewhat forced metaphor to describe what a bird does when it sings. From the point of view of analyzing the semantic gestalt, we may note that there is dissonance with the notion of throwing out unwanted stuff.

7. We all know … how a singing bird makes us feel and we can imagine how the bird feels as it **throws its song out** into the air. —(BNC) Julia Casterton, 1992. *Creative writing: A practical guide*.

Finally, 8 exemplifies an exploitation embedded within an exploitation. The scene is set with a conventional metaphor ('metamorphosis'), but then a new metaphor is introduced—resonating with but not actually realizing the conventional metaphors *thrown in at the deep end* (and perhaps also the conventional expressions *thrown into doubt* and *thrown into confusion*). Not only is the writer here comparing his younger self to an insect ('metamorphosis'), but also the deep end of the conventional swimming pool into which beginners are thrown has been metamorphosed into a jungle—a jungle of jargon (note the alliteration). This plethora of mixed metaphors and stylistic devices may offend stylistic purists and pedants, but that is a matter of taste. It is hard to sustain the argument that the intended meaning of the text is obscured or diminished by them, and some readers may indeed feel that the meaning is enhanced by them.

8. It was to take me some time longer to undergo the metamorphosis from a 'teacher' to a 'lecturer'. I was **thrown into a jungle** of new jargon. The language of special education had long been tucked under my belt, but now I was faced with filling in timetables with terms such as 'DD' time—departmental duties, to the uninitiated—in other words, time when I was not actually in direct teaching contact with students.
   —(BNC) Tony Booth et al. 1992. *Policies for diversity in education*.

4.3 Find the pattern: what is a pattern?

It should be clear by now that, in my view, the sort of thing that needs to be said in order to report corpus evidence accurately differs greatly from what is said in currently available dictionaries. In the first place, dictionaries must in future focus more on reporting conventional patterns of usage. In the second place, lexical analysts also need an apparatus for recognizing and classifying different sorts of exploitations, if only so as to know what to leave out, thus avoiding errors such as reporting solemnly, as one best-selling American dictionary does, that the word *newspaper*, in addition to being a noun meaning "a paper that is printed and distributed usu. daily or weekly and that contains news, articles of opinion, features, and advertising", is also a verb meaning "to do newspaper work".

In order to report patterns of conventional usage of words accurately, e-lexicographers must aim to analyse corpora and map meanings onto the patterns found. To do this, they need access to at least the following kinds of tools for processing corpus data:

- A part-of-speech analyser
- A sentence parser
- A system for organizing lexical items in the co-text around a target word into lexical sets, in most if not all cases according to some unifying semantic type or other semantic feature.

At the heart of each pattern lies a word. Patterns represent a combination of valency and collocational preferences of the target word.

4.4 Are word senses mutually exclusive?

Existing dictionaries postulate a comparatively small number of senses for each word, but say little about how each sense is realized in ordinary usage. As computational linguists have discovered belatedly and to their chagrin and cost (see, for example, Ide and Wilks 2006), not only is there very little indication in dictionaries of how one sense is to be distinguished from another, but also, to make matters worse, the senses in traditional dictionaries are not mutually exclusive. For example, two of the senses listed by OALDCE for the verb *pour* are:

> 1. [VN, usually + *adv/prep*] to make a liquid or other substance flow from a container in a continuous stream, especially by holding the container at an angle.

and

> 3. ~ (**sth**) (**out**) to serve a drink by letting it flow from a container into a cup or glass.

There almost total semantic overlap here. (What is a drink if it is not a liquid? What is the difference between “to serve by letting it flow” and “to make a liquid … flow”?) Also, there are a number of false implications lying in wait to ambush a naïve user or computer program, especially one that is looking for disambiguation criteria. The grammatical information given with sense 3 might look as if its purpose is disambiguation, but it is not, because:

- The word ‘**out**’ in sense 3 is a completive-intensive, not a disambiguator. The round brackets indicate that it is optional.
- ‘**out**’ in sense 3 is nothing more than a realization of the adv/prep mentioned in 1. It does not have a semantically distinctive function.

This is not a criticism of the OALDCE entry or of similar treatments of this word in other dictionaries. Rather, it is partly a criticism of naive expectations among dictionary users and partly a criticism of the theoretical foundations and assumptions underlying pre-corpus dictionaries. The entry for *pour* in OADLCE is not an isolated example. Many other entries in many other dictionaries exhibit similar undisambiguated semantic overlap. For a human reader, such redundancy may have a reinforcing rather than a confusing effect. However, for hard-nosed linguistic theory and for computational and artificial-intelligence, this sort of misinterpretation of the semantic redundancy in dictionaries can have consequences that are fatal to language understanding and practical applications.

When we turn to corpus evidence and examine it carefully, we find that disambiguation is indeed possible, but not word by word. Rather, it has to be undertaken phrase by phrase. Moreover, it turns out to be necessary to abandon the comfortable expectation that all possible uses of each word can be covered in a dictionary entry. Words have innumerable rare and improbable but possible uses, for language is dynamic: a design feature of the lexicon of a natural language is that it is made to be used creatively and innovatively as well as conventionally. The only realistic possibility for lexicographers is to aim to cover all probable uses of each word—that is, all normal, conventional uses. Each word in a language is associated with a small number of recurrent phraseological patterns. In most but not quite all cases, a unique sense can be associated with each pattern. A phraseological pattern consists of a mixture of valency and collocational preferences (see Hanks and Pustejovsky 2005).

Let us look at the entry for *pour* in the very first dictionary of English that took phraseology and collocations seriously, namely Cobuild (1987). It, too, made a distinction between pouring liquid out of a container and pouring a drink, but with this difference: typically, pouring a drink takes a benefactive argument. You pour someone a drink, but you don't pour someone petrol. Here is what Cobuild says:

> 1. If you **pour** a liquid or other substance, you cause it to flow out of a container by holding the container at a particular angle.
>
> 2. If you **pour** *someone* a drink, you fill a cup or glass with the drink so that they can drink it. [My italics]

Contrasts between word senses as presented in dictionaries and pattern senses as presented in corpora offer rich opportunities for future study. So far, neither dictionaries nor grammars have succeeded in defining or demonstrating systematically the associations between meaning and phraseological patterns at a sufficiently delicate level for reliable disambiguation of words in free text. This could be a major goal for future e-lexicography.

Accurate description of lexical patterns requires analysis not only of the syntactic structures (valencies) in which each word participates, but also analysis of collocational preferences within such structures.

More fundamentally, lexicography is now in a position to spearhead (or, if you prefer a different metaphor, to provide the foundations for) radical new approaches to the theoretical understanding of meaning in language. It is not entirely clear why 20th-century linguistics (in the English-speaking world) found it necessary to place so much emphasis on syntax, while having so little to say about lexis and meaning. What is clear now, in the age of electronic text processing, is that traditional assumptions about the nature of meaning and its relation to syntax are due for an overhaul and that lexicography and lexical studies are in a position to lead the way.

5. Is Wiktionary the right model for electronic dictionaries of the future?

What will major innovative dictionaries of the future be like? We don't know. Printed books are likely to remain extremely conservative and command little or no serious investment and hence bring little or no serious innovation. Future large-scale new dictionaries are likely to be electronic products, but a stable business model (or academic funding model) that would justify large-scale investment in such innovations has not yet emerged.

Some people argue that all information should be free, and point to the great success of Wikipedia as a free information source. The success of Wikipedia is undeniable. However, the success of its companion project, Wiktionary, "a collaborative project for creating a free lexical database in every language, complete with meanings, etymologies, and pronunciations", is less obvious. The contrast between Wikipedia and Wiktionary deserves a moment's consideration: it highlights the difference between a dictionary and an encyclopedia. If a reader wants expert information on some subject—let us say the reason why gold is valuable, or the characteristics of the Tocharian languages—he or she needs information from an expert. In the vast majority of Wikipedia articles, such expertise will be found, amply confirmed by other experts in the community, in accordance with what Putnam (1974) called 'the division of linguistic labour': you and I may not be able to distinguish *gold* from iron pyrites (called 'fool's gold') or other metals, but we rely on there being someone in the English-speaking world who can.

Part of the genius of Jimmy Wales and Larry Sanger, co-founders of Wikipedia, was to recognize that there are a) enough people in the world ready and willing to write and publish well-informed, accurate, and reliable articles about almost every topic under the sun without pay, and b) enough people in the world to spot poor or unreliable articles and be motivated to complain and even to provide something better: Wikipedia is truly a worldwide collective social endeavour. Part of their naivety was not to allow for the possibility that pranksters would try to slip in false, malicious, and/or damaging articles, as happened in the case of the Seigenthaler incident (2005: http://en.wikipedia.org/wiki/Seigenthaler_incident), as a result of which Wikipedia introduced new control and vetting procedures. Such incidents and shortcomings are now rare and Wikipedia has procedures in place for correcting them immediately if spotted.

So how does the Wiktionary enterprise match up to its encyclopedic brother? The avowed aim is "to include not only the definition of a word, but also enough information to really understand it" (http://en.wiktionary.org/wiki/Wiktionary:Main_Page). This laudable aim is inspiring, but at present it is not achieved. In the English Wiktionary, the etymologies are taken from or based on those in older dictionaries; as are definitions, which are extremely old-fashioned and derivative, taking no account of recent research in either cognitive linguistics or corpus linguistics.

Two brief examples may be given. If the aim is to give enough information for people to "really understand" the meanings of words, then some account must be given of, among other things, the research that has shown that the conventional phraseology associated with each word helps to determine its meaning and the research that has shown that much meaning in everyday language is metaphorical in nature.

Let us look in a little more detail at each of these two points in turn, with examples. As mentioned above, the CPA research project (http://nlp.fi.muni.cz/projects/cpa/) has shown that the meaning of a verb is closely allied to the semantic types of its arguments. Thus, the following (from BNC) are examples of the most normal uses of the verb *admit*:

1. At least three people were admitted to hospital.
2. Julie Smith was admitted for an emergency appendicectomy.
3. John was admitted into a local residential home.
4. The children were eventually admitted into care as a result of neglect.
5. Namibia was formally admitted to the UN as the organization's 160th member on April 23, 1990.

The collocations and the passive voice in 1-5 clearly distinguish this meaning of *admit* from other meanings of the same verb such as 'say reluctantly'. How does this work? First, note that we are talking here about normal, typical usage, not all possible uses. This is very important. The verb in this sense is normally passive, while in the 'say reluctantly' sense it is normally active—although the converse is also possible (as in *the hospital admitted three people; negligence was admitted*). A painful lesson for linguistics of the past thirty years (though some people are reluctant to admit it) is that all linguistic analysis and especially lexical analysis must be conducted in terms of probabilities, not in terms of necessary and sufficient conditions. In 1-4 the combination of a human subject with the expressions 'to hospital', 'into a residential home', 'into care', and 'for an appendicectomy' select the sense 'be brought officially to a place where one can be looked after or treated medically, according to need'. Additionally, these collocations assign to the human subject of the passive verb the role of being a person who is suffering or judged to be in need.

In 5, the combination of 'Namibia' with 'to the UN' activates a related but slightly different sense, involving becoming a member of an organization, rather than being taken to a place in order to be looked after. Each of the words and phrases cited in the preceding paragraph forms part of a contextually relevant lexical set. Paradigmatic lexical sets of this kind have (or may have) a very large number—indeed, an open-ended number—of words and phrases as members, but they are united by certain shared semantic features. Part of the art of electronic lexicography in the future will consist of selecting from a corpus typical examples of such lexical sets and summarizing their semantic structure in different contexts. This is not a simple task: for example, in the above examples 'hospital' and 'residential home' can be classified easily enough as locations, with the added proviso that these are locations in which care is given. However, the word 'care' itself does not denote a location. Nevertheless, we can be certain that if a text says that person is 'admitted into care', they are admitted to such a location. Thus, to borrow a term from Pustejovsky's Generative Lexicon theory (1995), *care* in this context is <u>coerced</u> into implying a location or, more specifically, a residential home. Notice, too, that conventional phraseology of this kind fills in all sorts of other gaps that are not explicitly stated but subliminally present: we know or can surmise that the people in 1 were injured or ill, that Julie in 2 went to a hospital, that John in 3 was a child or disabled person in need of care, and that in 4 the children were taken to a care home, although these facts are not explicitly stated. This kind of implicature is also one of the fundamental insights of Fillmore's Frame Semantics (1982, 2006).

Now compare how English Wiktionary (accessed 27 March 2011) defines and exemplifies this sense of the word:

> To allow to enter; to grant entrance, whether into a place, or into the mind, or consideration; to receive; to take.
>
> > *A ticket admits one into a playhouse.*
> >
> > *They were admitted into his house.*
> >
> > *to admit a serious thought into the mind*
> >
> > *to admit evidence in the trial of a cause*

The Wiktionary definition is not wrong, but it is stilted and archaic in wording (note, for example, the old-fashioned uses of 'grant' and 'whether') and it does not record that in this sense the verb is usually passive. It does not do a good job of explaining the meaning, which, in modern English, has more to do with activating an administrative procedure than with "allowing", "granting", "receiving", or "taking". The location to which admission is granted, as we have seen, is generally an institution of some kind, rather than "the mind, or consideration". (I could go on.) Wikitionary's examples do not illustrate the normal phraseology in which the verb is used: instead, they seem to be intended to illustrate extreme possibilities of usage. The examples were, needless to say, are not corpus-based; they were invented by a lexicographer, either recently or a hundred years ago. If they seem stilted and unnatural, it is because most human beings are, strangely, not very good at reporting or inventing examples of their own normal, everyday linguistic behaviour. The human mind, when pressed for an example, seems to reach unerringly for a boundary case, even at the expense of idiomaticity and naturalness, rather than for a central and typical, normal example.

If we turn now from verbs to Wikitionary's treatment of concrete nouns, we can illustrate our second point, namely how a traditional approach to definition fails to record essential components of lexical meaning that provide the foundations of everyday metaphorical exploitations of meaning. A very simple example is the conventional simile *treat someone like a dog*. This means 'to treat someone badly', despite the fact that in most English-speaking cultures dogs are horribly pampered and well treated. It is necessary to distinguish between creative figurative language—genuine exploitations of conventional norms—and conventional figurative expressions. The latter type of figurative language deserve to be recorded in dictionaries, though at present this is not done systematically in any dictionary. A vast mass of research over the past thirty years has revealed that metaphor plays a central role in everyday linguistic meaning. See for example Lakoff and Johnson (1980), Glucksberg (2001), Giora (2003), Bowdle and Gentner (2005), and collections of papers such as those in Stefanowitsch and Gries (eds, 2006), Gibbs (ed., 2008), and Hanks and Giora (2011 [In Press]).

Conventional figurative exploitations of the meaning of *dog* are too many and complex to allow full discussion here. Let us instead take a simpler word: *elephant*. The Wiktionary definition of *elephant* (accessed 29 March 2011) reads as follows:

1. A mammal of the order *Proboscidea*, having a trunk, and two large ivory tusks jutting from the upper jaw.
2. (*figuratively*) Anything huge and ponderous.
3. (*paper, printing*) A printing-paper size measuring 30 inches x 22 inches.
4. (*UK, childish*) used when counting to add length. *Let's play hide and seek. I'll count. One elephant, two elephant, three elephant...*

Leaving aside senses 3 and 4, we may ask whether definitions 1 and 2 give a satisfactory account of the meaning. They do not. In the first place, sense 1 fails to say that elephants are large, a fact often exploited metaphorically. It is not the case that elephants are necessarily large, but they are typically large. This, too, is an important point. The discovery of dwarf elephants in Borneo and skeletons of an extinct species of dwarf elephants in Crete must not be allowed to inhibit the lexicographer from saying that elephants are typically large. Only this will enable interpretation of examples such as 6 and 7 below. It also needs to be said that elephants have a proverbially good memory (8 and 9), that bull elephants are assertive (10), and that they make an extremely loud noise called trumpeting (11).

6. So what I'm actually saying is that I'm making my objective an elephant, it's too large. —(BNC) *spoken corpus; staff training session.*
7. Och, I don't want a stranger to think that I'm built like an elephant. —(BNC) *spoken corpus; unscripted conversation.*
8. "You've got the memory of an elephant, you're probably the cleverest girl in class and you can't read." —(BNC) Celia Brayfield, 1990. *The Prince*.
9. But the odd rumour has gone round that Six has been operating someone big, someone quite high up in the KGB—someone with an elephant 's memory who might be about to finger Mills once and for all. —(BNC) Trevor Barnes, 1991. *A Midsummer Killing*.
10. He turned round to gaze at Cord Dillon, Deputy Director of the CIA. "A rough diamond," Paula called him. "The manners of a bull elephant," was Monica's elegant description. —(BNC) Colin Forbes, 1991. *Whirlpool.*
11. Then someone asked me where the station was, and she was deaf, and I had to trumpet like an elephant for about ten minutes. —(BNC) Mary Gervaise, 1983. *The distance enchanted*.

I have illustrated just two of the many kind of improvements that could be made to a dictionary such as English Wiktionary using corpus evidence. The essential message here is that, as in many traditional dictionaries, the definitions may succeed in defining, but they do not do a very good job of explaining. Is it really of any help to anyone (except, perhaps, a taxonomic zoologist) to be told that an elephant is "a mammal of the order *Proboscidea*" or that an elephant seal is "a large marine mammal of the genus *Mirounga*, which is the largest of the pinnipeds"?

Does all this rule out Wikitionary as a model for electronic lexicography? No, absolutely not. There are many positive things to be said. In the first place, Wiktionary shows how imaginative use can be made of multimedia hypertext links such as audio links to the pronunciation of the word in different standard accents of English (American and British), pictures of elephants, and text links to related terms such as *elephant seal, elephant shrew* (so called because of its long nose), *white elephant,* and *pink elephant*. Some links lead to the encyclopedic rather than the lexicographic components of the Wikimedia complex, which seems just right for a natural-kind term such as *elephant*. No doubt it would be technically straightforward enough to include film clips of typical elephant behaviour, including the sound of elephants trumpeting. Perhaps the technology is not far away by which we shall be able to sit at our computer and touch a simulation of an elephant's skin or smell a bull elephant in musth.

It is a cardinal point of principle for all Wikimedia that the information supplied should be freely available to everybody. This does not mean, however, that there is an absence of control. I noticed that there was no entry in Wikipedia for the term *rogue elephant*, so, as an experiment, I added a definition for the figurative sense, "someone or something that is large, dangerous, and unpredictable." Within minutes, my tiny contribution had been placed within a template for noun entries, and someone (presumably at Wiktionary central) had added the literal sense, "A solitary, old, male elephant that has become dangerously and unpredictably violent", together with a cross-reference to a term with similar meaning, *loose cannon*. This is very impressive.

Similar controls are evidently [CHECK?] also in place to prevent use of Wikitionary for propaganda purposes, for example by religious groups such as Scientologists or lawyers representing commercial conglomerates such as the Edgar Rice Burroughs Foundation, both of whom have put pressure in the past on dictionary publishers to amend definitions to show their products ('Scientology' and 'Tarzan' theme parks respectively) in a favourable light.

The hypertext structure of Wikimedia, and in particular Wiktionary, is eminently suitable as a model for the electronic dictionary of the future. What is needed is some way of ensuring that definitions are properly supported by links to corpus evidence, including evidence for the ways in which word meanings are exploited in metaphors and other ways. This will entail that almost every definition of every content word must be radically re-examined in the light of corpus evidence, in the way suggested above. Such a re-examination needs to be conducted systematically and professionally. Our sympathies may be with an anarcho-syndicalist approach, but it is hard to imagine how a radical new approach to defining verbs or natural-kind terms could be carried out systematically by enthusiasts and volunteers. Nevertheless, the overall aim must remain "to include not only the definition of a word, but also enough information to really understand it".

6. Conclusions

Lexicography is in a state of transition at the present time, between the technology of the printed word and the bound book that has served us so well for 500 years and the technology of the Internet and the electronic product; also between exploded Leibnizian assumptions about the relationship between words and concepts and newer theories of prototypes and stereotypes based on the work of philosophers such as Wittgenstein, Grice, and Putnam and cognitive scientists such as Rosch, Lakoff, Gentner, and Glucksberg.

It is too early to say what form innovative dictionaries of the future will take. Perhaps the Wiktionary model can be adapted, or perhaps an entirely new business model will be developed by an enterprising electronic publisher. One thing seems certain, however: all serious future lexicography will be corpus-driven, not merely a matter of guesswork based on speculation.

References

Atkins, B. T. Sue, and Beth Levin. 1991. 'Admitting impediments'. In U. Zernik (ed.), *Lexical Acquisition: Using On-Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates.

Bowdle, Brian, & Dedre Gentner (2005). 'The career of metaphor'. In *Psychological Review*, 112 (1).

Church, Kenneth W., and Patrick Hanks. 1989. 'Word Association Norms, Mutual Information, and Lexicography'. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 26-29 June 1989*. University of British Columbia. Revised version published in *Computational Linguistics*, 16 (1). Reprinted in Fontenelle (ed., 2008) and Hanks (ed., 2008a).

Fillmore, Charles J. 1982. 'Frame semantics'. In *Linguistics in the Morning Calm*. Hanshin Publishing Co.

Fillmore, Charles J. 2006. 'Frame Semantics'. In K. Brown (ed.). *Encyclopedia of Language and Linguistics, 2nd edition*. Elsevier.

Fillmore, Charles J., and B. T. S. Atkins. 1992. 'Towards a frame-based lexicon: the semantics of ***risk*** and its neighbors'. In Adrienne Lehrer and Eva Feder Kittay (eds., 1992): *Frames, Fields, and Contrasts.* Lawrence Erlbaum Associates.

Firth, J. R. 1950. 'Personality and language in society'. In *The Sociological Review,* xlii. Reprinted 1957 in *Papers in Linguistics 1934-1951*. Oxford University Press.

Firth, J. R. 1957a. 'Modes of Meaning'. In *Papers in Linguistics 1934-1951*. Oxford University Press.

Firth, J. R. 1957b. 'A synopsis of linguistic theory 1930-1955'. In *Studies in Linguistic Analysis*. Philological Society, Oxford. Reprinted in F. R. Palmer (ed.; 1968), *Selected Papers of J. R. Firth*. Longman.

Gibbs, Raymond W., Jr. (ed.) 2008. *The Cambridge Handbook of Metaphor and Thought*. Cambridge University Press.

Giora, Rachel. 2003. *On our mind: Salience, context, and figurative language*. Oxford University Press.

Glucksberg, Sam. 2001. *Understanding Figurative Language.* Oxford University Press.

Hanks, Patrick (ed.). 2008a. *Lexicology: Critical Concepts in Linguistics.* 6 volumes. Routledge.

Hanks, Patrick. 2008b. 'The lexicographical legacy of John Sinclair'. In *International Journal of Lexicography (2008),* 21 (3).

Hanks, Patrick. 2009. 'The Impact of Corpora on Dictionaries' in Paul Baker (ed.), *Contemporary Corpus Linguistics.* Continuum.

Hanks, Patrick. 2010. 'Terminology, Phraseology, and Lexicography.' In A. Dykstra et al. (eds), *Euralex Proceedings*. Leeuwarden: Frisian Institute.

Hanks, Patrick. [in press (2011a)]. 'Wie man aus Wörtern Bedeutungen macht: semantische Typen treffen syntaktische Dependenzen'. In Proceedings of the 2010 Jahrestagung of the Institut für Deutsche Sprache, Mannheim.

Hanks, Patrick [in press (2011b)]. 'Representing the unrepresentable: Dictionaries, documents, and meaning'. In Pier Marco Bertinetto, Valentina Bambini, Irene Ricci, and collaborators (eds.), *Linguaggio e cervello – Semantica / Language and the brain – Semantics*. Atti del XLII Convegno della Società di Linguistica Italiana (Pisa, Scuola Normale Superiore, 25-27 settembre 2008). Roma, Bulzoni.

Hanks, Patrick [Forthcoming (2012)]. *Lexical Analysis: Norms and Exploitations.* MIT Press.

Hanks, Patrick, and Rachel Giora (eds). [In press (2011)]. *Metaphor and Figurative Language: Critical Concepts.* 6 volumes. Routledge.

Hanks, Patrick, and James Pustejovsky. 2005. 'A pattern dictionary for natural language processing'. In *Revue Française de Langue Appliquée*, 10:2.

Ide, Nancy, and Yorick Wilks. 2006. 'Making Sense About Sense'. In E. Agirre and Philip Edmonds (eds.), *Word Sense Disambiguation: Algorithms and Applications.* Springer.

Kilgarriff, Adam. 2005. 'Language is never ever, ever, ever random'. In: *Corpus Linguistics and Linguistic Theory* 1 (2).

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press.

Sinclair, John. 1966. 'Beginning the study of lexis'. In C.E. Bazell, J.C. Catford, M.A.K. Halliday, and R.H. Robins (eds.), *In Memory of J. R. Firth*. Longman. Reprinted in Reprinted in Hanks (ed. 2008a), vol. 4.

Sinclair, John. 1984. 'Naturalness in language'. In J. Aarts and W. Meijs (eds.), *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Rodopi.

Sinclair, John. 1987. 'The Nature of the evidence' In: J. Sinclair (ed.), *Looking Up*. Collins Publishers.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press,

Sinclair, John. 2010. 'Defining the definiendum' in G.-M. de Schryver (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis.* Kampala and Ghent: Menha Publishers.

Web sites

Pattern Dictionary of English Verbs at http://nlp.fi.muni.cz/projects/cpa/

FrameNet: at http://framenet.icsi.berkeley.edu/