Automating the creation of dictionaries: where will it all end?
Michael Rundell and Adam Kilgarriff

**Abstract**

The relationship between dictionaries and computers goes back around 50 years. But for most of that period, technology's main contributions were to facilitate the capture and manipulation of dictionary text, and to provide lexicographers with greatly improved linguistic evidence. Working with computers and corpora had become routine by the mid-1990s, but there was no real sense of lexicography being automated. In this article we review developments in the period since 1997, showing how some of the key lexicographic tasks are beginning to be transferred, to a significant degree, from humans to machines. A recurrent theme is that automation not only saves effort but often leads to a more reliable and systematic description of a language. We close by speculating on how this process will develop in years to come.

## 1. Introduction

This paper describes the process by which – over a period of 50 years or so – several important aspects of dictionary creation have been gradually transferred from human editors to computers. We begin by looking at the early impact of computer technology, up to and including the groundbreaking COBUILD project of the 1980s. The period that immediately followed saw major advances in the areas of corpus building and corpus software development, and the first dedicated dictionary writing systems began to appear. These changes – important though they were – did not significantly advance the process of automation. Our main focus is on the period from the late 1990s to the present. We show how a number of lexicographic tasks, ranging from corpus creation to example writing, have been automated to varying degrees. We then look at several areas where further automation is achievable and indeed already being planned. Finally, we speculate on how much further this process might have to run, and on the implications for dictionaries, dictionary-users, and dictionary-makers.

## 2. Computers meet lexicography: from the 1960s to the 1990s

The great dictionaries of the 18th and 19th centuries were created using basic technologies: pen, paper, and index cards for the lexicography, hot metal for the typesetting and printing. In the English-speaking world, the principle that a dictionary should be founded on objective language data was established by Samuel Johnson, and applied on a much larger scale by James Murray and his collaborators on the Oxford English Dictionary (OED, Murray et al. 1928). The task of collecting source material – citations extracted from texts – was immensely laborious. Johnson employed half a dozen assistants to transcribe illustrative sentences which he had identified in the course of his extensive reading, while the OED's 'corpus' – running into several million handwritten 'slips' – was collected over several decades by an army of volunteer readers. And this

was only the first stage in the dictionary-making process. In all of its components, the job of compiling a dictionary was extraordinarily labour-intensive. Johnson's references to 'drudgery' are well-known, but Murray's letters testify even more eloquently to the stress, exasperation, exhaustion and despair which haunted his life as the OED was painstakingly assembled (Murray 1979, esp. Ch XI).

It was Laurence Urdang – as Editor of the *Random House Dictionary of the English Language* (Stein and Urdang 1966) – who first saw the potential of computers to facilitate and rationalize the capture, storage and manipulation of dictionary text.[1] From this point, the idea of the dictionary as a database, in which each of the components of an entry has its own distinct field, became firmly established. An early benefit of this approach was that cross-references could be checked more systematically: the computer generated an error report of any cross-references that did not match up, and errors would then be dealt with manually. An extremely dull task was thus transferred from humans to computers, but with the added benefit that the computers made a much better job of it. And when learner's dictionaries began to control the language of definitions by using a limited defining vocabulary (DV), similar methods could be used to ensure that proscribed words were kept out. In a further development, the first edition of the *Longman Dictionary of Contemporary English* (LDOCE1, 1978) included some categories of data (notably a complex system of semantic coding) which were never intended to appear in the dictionary itself. In projects like these, the initial text-compilation process remained largely unchanged, but subsequent editing was typically done on pages created by line printers, with the revisions keyed into the database by technicians.

### *2.1 Year Zero: the COBUILD project*

Some time around 1981 marks Year Zero for modern lexicography. The COBUILD project brought many innovations in lexicographic practices and editorial styles (as described in Sinclair 1987), but our focus here is on the impact of technology, and its potential to take on some of the tasks traditionally performed by humans. Computers were central to the COBUILD approach from the start. Like the visible tip of an iceberg, the eventual dictionary would be derived from a more extensive database, and lexicographers created their entries using an array of coloured slips to record information of different types (Krishnamurthy 1987). Every linguistic fact the lexicographers identified would be supported by empirical evidence in the form of corpus extracts. For the first time, a large-scale description of English was created from scratch to reflect actual usage as illustrated in (what was then) a large and varied corpus of texts. The systematic application of this corpus-based methodology represents a paradigm shift in lexicography. What was revolutionary in 1981 is now, a generation later, the norm for any serious lexicographic enterprise. But from the point of view of the human-machine balance, COBUILD's advances were relatively modest. Corpus creation was still a laborious business. As the use of scanners supplemented keyboarding, data capture was somewhat less arduous than the methods available to Henry Kučera two decades earlier, when he used punched cards to turn a million words of text into the Brown Corpus (Kučera & Francis 1967). But like their predecessors at Brown, the COBUILD

**Komentář [MU1]:** Or corpus-based or corpus-informed ? Corpus-driven has other implications (see e.g. Tongini-Bonelli 2001).

developers were testing available technology to its limits, and building the corpus on which the dictionary would be based involved heroic efforts (Renouf 1987). As for the lexicographic team, few ever got their hands on a computer. Concordances were available in the form of microfiche printouts, and the fruits of their analysis were written in longhand – the slips then being handed over to a separate team of computer officers responsible for data-entry.

*2.2 The 80s and 90s*

The fifteen years or so that followed saw quite rapid technical advances. Computers moved from being large and expensive machines available only to specialists, to become everyday objects to be found on most desks in the developed world. This has brought vast changes to many aspects of our lives. During this period, corpora became larger by an order of magnitude, and improved corpus-query systems (CQS) enabled lexicographers to search the data more efficiently. The constituent texts of a corpus were now routinely annotated in various ways. Forms of annotation included tokenization, lemmatization, and part-of-speech tagging (see Grefenstette 1998: 28-34 and Atkins & Rundell 2008: 84-92 for summaries), and this allowed more sophisticated, better-targeted searches. From the beginning of the 1990s, it became normal for lexicographers to work on their own computers rather than depending on technical staff for data-entry, and the first generation of dedicated dictionary-writing systems (DWS) were created.

By the late 1990s, the use of computers in data analysis and dictionary compilation was standard practice (at least for English). But to what extent was lexicography 'automated' at this point? Corpus creation remained a resource-intensive business. Corpus analysis was easier and faster, but lexicographers found themselves handling far more data. From the point of view of producing more reliable dictionary entries, access to higher volumes of data was a good thing. But scanning several thousand concordance lines for a word of medium frequency (within the time constraints of a typical dictionary project) is a demanding task – in a sense, a new form of drudgery for the lexicographer.

On the entry-writing front, the new DWS made life somewhat easier. When we use this kind of software, the overall shape of an entry is controlled by a 'dictionary grammar'. This in turn implements the decisions made in the dictionary's style guide about how the many varieties of lexical facts are to be classified and presented. Data fields such as style labels, syntax codes, and part-of-speech markers have a closed set of possible contents which can be presented to the compiler in drop-down lists. Lexicographers no longer have to remember whether a particular feature should appear in bold or italics, whether a colloquial usage is labelled 'inf', 'infml' or '*informal*', and so on. In areas like these, human error is to a large extent engineered out of the writing process. A good DWS also facilitates the job of editing. For example, an editor will often want to restructure long entries, changing the ordering or nesting of senses and other units. This is a hard intellectual task, but the DWS can at least make it a technically easy one.

**Komentář [MU2]:** facts?

Meanwhile, some essential but routine checks – cross-reference validation, defining vocabulary compliance, and so on – are now fully automated, taking place *at the point of compilation* with little or no human intervention.

With more linguistic data at their disposal and better software to exploit it, and with compilation programs which strangle some classes of error at birth, support the editing process, and quietly handle a range of routine checks, lexicographers now had the tools to produce better dictionaries: dictionaries which gave a more accurate account of how words are used, and presented it with a degree of consistency which was hard to achieve in the pre-computer age.

Whether this makes life easier for lexicographers is another question. Delegating low-level operations to computers is clearly a benefit for all concerned. The computers do the things they are good at (and do them more efficiently than humans), while the lexicographers are relieved of the more tedious, undemanding tasks and thus free to focus on the harder, more creative aspects of dictionary-writing. But the effect of these advances is limited. The core tasks of producing a dictionary still depend almost entirely on human effort, and there is no sense, at this point, of lexicography being automated.

## 3. From 1997 to the present

What we describe above represents the state of the art in the late 1990s. For present purposes, we will take as our baseline the year 1997, which is when planning began for a new, from-scratch learner's dictionary.

If the big change to the context of working life in the 80s and 90s was that most of us (in lexicography and everywhere else) got a computer, the big change in the current period is that the computer got connected to the Internet.

When work started on the *Macmillan English Dictionary for Advanced Learners* (Rundell, ed., 2001), we had the advantage of entering the field at a point when the corpus-based methodology was well-established, and the developments described above were in place. But we faced the challenge of entering a mature market in which several high-quality dictionaries were already competing for the attention of language learners and their teachers. It was clear that any new contender could only make a mark by doing the basic things well, and by doing new things which had not been attempted before but which would meet known user needs. It was equally clear that computational methods would play a key part in delivering the desired innovations.

The rest of this paper reviews developments in the period from 1997 to the present, and discusses further advances that are still at the planning stage. The work we describe represents a collaboration between a lexicographer and a computational linguist (the authors), and shows how the job of dictionary-makers has been supported by, and in some cases replaced by, computational techniques which originate from research in the field of natural language processing (NLP). We will conclude with some speculations on

the direction of this trajectory: is the end point a fully-automated dictionary? does it even make sense to think in terms of an 'end-point'?

First, it will be helpful to give a brief inventory of the main tasks involved in creating a dictionary, so that we can assess how far we have progressed along the road to automation. They are:

- corpus creation
- headword list development
- analysis of the corpus:
  - to discover word senses and other lexical units (fixed phrases, phrasal verbs, compounds, etc.)
  - to identify the salient features of each of these lexical units
    - (1) their syntactic behaviour
    - (2) the collocations they participate in
    - (3) their colligational preferences
    - (4) any preferences they have for particular text-types or domains
- providing definitions (or translations) at relevant points
- exemplifying relevant features with material gleaned from the corpus
- editing compiled text in order to control quality and ensure consistent adherence to agreed style policies

We look at all of these, some in more detail than others.

*3.1 Corpus creation*

For people in the dictionary business, one of the most striking developments of the 21st century is the 'web corpus'. Corpora are now routinely assembled using texts from the Internet and this has had a number of consequences. First, the curse of data-sparseness, which has dogged lexicography from Johnson's time onwards, has become a thing of the past.[2] The COBUILD corpus of the 1980s – an order or magnitude larger than Brown – sought to provide enough data for a reliable account of mainstream English, but its creators were only too aware of its limitations.[3] The British National Corpus (BNC) – larger by another order of magnitude – was another attempt to address the issue.

As new technologies have arisen to facilitate corpus creation from the web, it has become possible to create register-diverse corpora running into billions of words. Software tools such as WebBootCat (Baroni & Bernardini 2004, Baroni et al. 2006) provide a one-stop operation in which texts are selected according to user-defined parameters, 'cleaned up', and linguistically annotated. The timescale for creating a large lexicographic corpus has been reduced from years to weeks, and for a small corpus in a specialised domain, from months to minutes. Texts on the web are, by definition, already in digital form. The overall effect is to drastically reduce both the human effort involved in corpus creation and the 'entry fee' to corpus lexicography.[4] Thus the process of collecting the raw data that will form the basis of a dictionary has to a large extent been automated.

**Komentář [MU3]:** Is the BANTU corpus also web-based ? AK:YES

Inevitably there are downsides. The granularity of smaller corpora (in terms of the balance of texts, the level of detail in document headers, and the delicacy of annotation) cannot be fully replicated in corpora of several billion words. While for some types of user (e.g. grammarians or sociolinguists) this will sometimes limit the usefulness of the corpus, for lexicographers working on general-purpose dictionaries, the benefits of abundant data outweigh most of the perceived disadvantages of web corpora. There were good reasons why the million-word Brown Corpus of 1962 was designed with such great care: a couple of 'rogue' texts could have had a disruptive effect on the statistics. In a billion-word corpus the occasional outlier will not compromise the overall picture. We now simply aim to ensure that the major text-types are all well represented.

Concerns about the diversity of text-types available on the web have proved largely unfounded. Comparisons of web-derived corpora against benchmark collections like the BNC have produced encouraging results, suggesting that a well-designed web corpus can provide reliable language data (Sharoff 2006, Baroni et al. 2009).[5]

### *3.2 Headword lists*

Building a headword list is the most obvious way to use a corpus for making a dictionary. *Ceteris paribus,* if a dictionary is to have N words in it, they should be the N words from the top of the corpus frequency list.

#### *3.2.1 In search of the ideal corpus*

It is never as simple as this, mainly because the corpus is never good enough. It will contain noise and biases. The noise is always evident within the first few thousand words of all the corpus frequency lists that either of us has ever looked at. In the BNC, for example, a large amount of data from a journal on gastro-uterine diseases presents noise in the form of words like *mucosa* – a term much-discussed in these specific documents, but otherwise rare and not known to most speakers of English.[6] Bias in the spoken BNC is illustrated by the very high frequencies for words like *plan, elect, councillor*, *statutory* and *occupational*: the corpus contains a great deal of material from local government meetings, so the vocabulary of this area is well represented. Thus keyword lists of the BNC in contrast to other large, general corpora show these words as particularly BNC-flavoured. And unlike many of today's large corpora, the BNC contains, by design, a high proportion of fiction. Finally, if our dictionary is to cover the varieties of English used throughout the world, the BNC's exclusive focus on British English is another limitation.

If we turn to UKWaC (the UK 'Web as Corpus', Baroni et al. 2009), a web-sourced corpus of around 1.6 billion words, we find other forms of noise and bias. The corpus contains a certain amount of web spam. In particular, we have discovered that people advertising poker are skilled at producing vast quantities of 'word salad' which is not easily distinguished – using automatic routines – from *bona fide* English. Internet-related bias also shows up in the high frequencies for words like *browser* and *configure*. While noise is simply wrong, and its impact is progressively reduced as ongoing cleanups are

implemented, biases are more subtle in that they force questions about the sort of language to be covered in the dictionary, and in what proportions.[7]

*3.2.2 Multiwords*

English dictionaries have a range of entries for multiword items, typically including noun compounds (*credit crunch*, *disc jockey*), phrasal and prepositional verbs (*take after*, *set out*) and compound prepositions and conjunctions (*according to*, *in order to*). While corpus methods can straightforwardly find high-frequency single-word items and thereby provide a fair-quality first pass at a headword list for those items, they cannot do the same for multiword items. Lists of high-frequency word-pairs in any English corpus are dominated by items which do not merit dictionary entries: the string *of the* is usually top of any list of bigrams. We have several strategies here: one is to view multiword headwords as collocations (see discussion below) and to find multiword headwords when working through the alphabet looking at each headword in turn. Another, currently underway in the Kelly project (Kilgarriff 2010) is to explore lists of translations of single-word headwords for a number of other languages into English, and to find out what multiwords occur repeatedly.

*3.2.3 Lemmatization*

The words we find in texts are inflected forms; the words we put in a headword list are lemmas. So, to use a corpus list as a dictionary headword, we need to map inflected forms to lemmas: we need to lemmatize.

English is not a difficult language to lemmatize as no lemma has more than eight inflectional variants (*be*, *am*, *is*, *are*, *was*, *were*, *been*, *being*), most nouns have just two (*apple*, *apples)* and most verbs, just four (*invade*, *invades*, *invading*, *invaded*). Most other languages, of course, present a substantially greater challenge. Yet even for English, automatic lemmatization procedures are not without their problems. Consider the data in Table 1. To choose the correct rule we need an analysis of the orthography corresponding to phonological constraints on vowel type and consonant type, for both British and American English.[8]

**Komentář [MU4]:** Hyphen or no hyphen ?? (AK : yes, consitency is good;-) – no hyphens is better)

Table 1: *Complexity in verb lemmatisation rules for English*

| **lemma** | | ***-ed, -s* forms** | **Rule** | ***-ing* form** | **Rule** |
|---|---|---|---|---|---|
| *fix* | | *fixed, fixes* | delete *–ed, -es* | *fixing* | delete *–ing* |
| *care* | | *cared, cares* | delete *–d, -s* | *caring* | delete *-ing*, add *-e* |
| *hope* | | *hoped, hopes* | delete *–d, -s* | *hoping* | delete *-ing*, add *-e* |
| *hop* | | *hopped* | delete *–ed*, undouble consonant | *hopping* | delete *-ing*, undouble consonant |
| | | *hops* | delete *–s* | | |
| *fuse* | | *fused* | delete *–d* | *fusing* | delete *-ing*, add *-e* |
| *fuss* | | *fussed* | delete *–ed* | *fussing* | delete *–ing* |
| *bus* | AmE | *bussed, busses*?? | delete *–ed/-s*, undouble consonant | *bussing* | delete *-ing*, undouble consonant |
| | BrE | *bused, bused* | delete *–ed* | *busing* | delete *–ing* |

Even with state-of-the-art lemmatization for English, an automatically extracted lemma list will contain some errors.

These and other issues in relating corpus lists to dictionary headword lists are described in detail in Kilgarriff (1997).

*3.2.4 Practical solutions*

Building a headword list for a new dictionary (or revising one for an existing title) has never been an exact science, and little has been written about it. Headword lists are by their nature provisional: they evolve during a project and are only complete at the end. A good starting point is to have a clear idea of what your dictionary will be used for, and this is where the 'user profile' comes in. A user-profile "seeks to characterise the typical user of the dictionary, and the uses to which the dictionary is likely to be put" (Atkins & Rundell 2008: 28). This is a manual task, but it provides filters with which to sift computer-generated wordlists.

An approach which has been used with some success is to generate a wordlist which is (say) 20% larger than the list you want to end up with – thus, a list of 60,000 words for a dictionary of 50,000 – and then whittle it down to size taking account of the user profile. Then, if the longer list contains obsolescent terms which are used in 19th century literature, but the user profile specifies that uses are all engaged with the contemporary language, these items could safely be deleted. If the user profile included literary scholarship, they could not.

*3.2.5 New words*

As everyone involved in commercial lexicography knows, neologisms punch far above their weight. They might not be very important for an objective description of the language but they are loved by marketing teams and reviewers. New words and phrases often mark the only obvious change in a new edition of a dictionary, and dominate the press releases.

Mapping language change has long been a central concern of corpus linguists, and a longstanding vision is the 'monitor corpus', the moving corpus that lets the researcher explore language change objectively (Clear 1988, Janicivic & Walker 1997). The core method is to compare an older 'reference' corpus with an up-to-the-minute one to find words which are not already in the dictionary, and which are in the recent corpus but not in the older one. O'Donovan & O'Neill (2008) describe how this has been done at Chambers Harrap Publishers, and Fairon et al. (2008) describe a generic system in which users can specify the sources they wish to use and the terms they wish to trace.

**Komentář [MU5]:** 1988 – see list of references

The nature of the task is that the automatic process creates a list of candidates, and a lexicographer then goes through them to sort the wheat from the chaff. There is always far more chaff than wheat. The computational challenge is to cut out as much chaff as possible without losing the wheat – that is, the new words which the lexicography team have not yet logged but which should be included in the dictionary.

For many aspects of corpus processing, we can use statistics to distinguish signal from noise, on the basis that the phenomena we are interested in are common ones and occur repeatedly. But new words are usually rare, and by definition are not already known. Thus lemmatization is particularly challenging since the lemmatizer cannot make use of a list of known words. So for example, in one list we found the 'word' *authore*, an incorrect but understandable lemmatization of *authored,* past participle of the unknown verb *author*.

For new-word finding we will want to include items in a candidate list even though they occur just once or twice. Statistical filtering can therefore only be used minimally. We are exploring methods which require that a word that occurred once or twice in the old material occurs in at least three or four documents in the new material, to make its way onto the candidate list. We use some statistical modulation to capture new words which are taking off in the new period, as well as the items that simply have occurred where they never did before. Many items that occur in the new words list are simply typing errors. This is another reason why it is desirable to set a threshold higher than one in the new corpus.

We have found that almost all hyphenated words are chaff, and often relate to compounds which are already treated in the dictionary as 'solid' or as multiword items. English hyphenation rules are not fixed: most word pairs that we find hyphenated (*sand-box*) can also be found written as one word (*sandbox*), as two (*sand box*), or as both. With this in mind, to minimise chaff, we take all hyphenated forms and two- and three-word items in the dictionary and 'squeeze' them so that the one-word version is included in the list of already-known items, and we subsequently ignore all the hyphenated forms in the corpus list.

Prefixes and suffixes present a further set of items. Derivational affixes include both the more syntactic (*-ly*, *-ness*) and the more semantic (*-ish, geo-, eco-*).[9] Most are chaff: we do not want *plumply* or *ecobuddy* or *gangsterish* in the dictionary, because, even though they all have google counts in the thousands, they are not lexicalised and there is nothing to say about them beyond what there is to say about the lemma, the affix and the affixation rule. The ratio of wheat to chaff is low, but amongst the nonce productions there are some which are becoming established and should be considered for the dictionary. So we prefer to leave the nonce formations in place for the lexicographer to run their eye over.

For the longer term, the biggest challenge is acquiring corpora for the two time periods which are sufficiently large and sufficiently well-matched. If the new corpus is not big enough, the new words will simply be missed, while if the reference corpus is not big enough, the lists will be full of false positives. If the corpora are not well-matched but, for example, the new corpus contains a document on vulcanology and the reference corpus does not, the list will contain words which are specialist vocabulary rather than new, like *resistivity* and *tephrochronology.*

While vast quantities of data are available on the web, most of it does not come with reliable information on when the document was originally written. While we can say with confidence that a corpus collected from the web in 2009 represents, overall, a more recent phase of the language than one collected in 2008, when we move to words with small numbers of occurrences, we cannot trust that words from the 2009 corpus are from more recently-written documents than ones from the 2008 corpus.

Two text types where date-of-writing is usually available are newspapers and blogs. Both of these have the added advantage that they tend to be about current topics and are relatively likely to use new vocabulary. Our current efforts for new-word-detection involve large-scale gathering of one million words of newspapers and blogs per day. The collection started in early 2009 and we need to wait at least one year or possibly two before we can assess what it achieves. Over a shorter time span lists will be dominated by short-term items and items related to the time of year. It will take a longer view to support the automatic detection of new words which have become established and have earned their place in the dictionary.

*3.3 Collocation and word sketches*

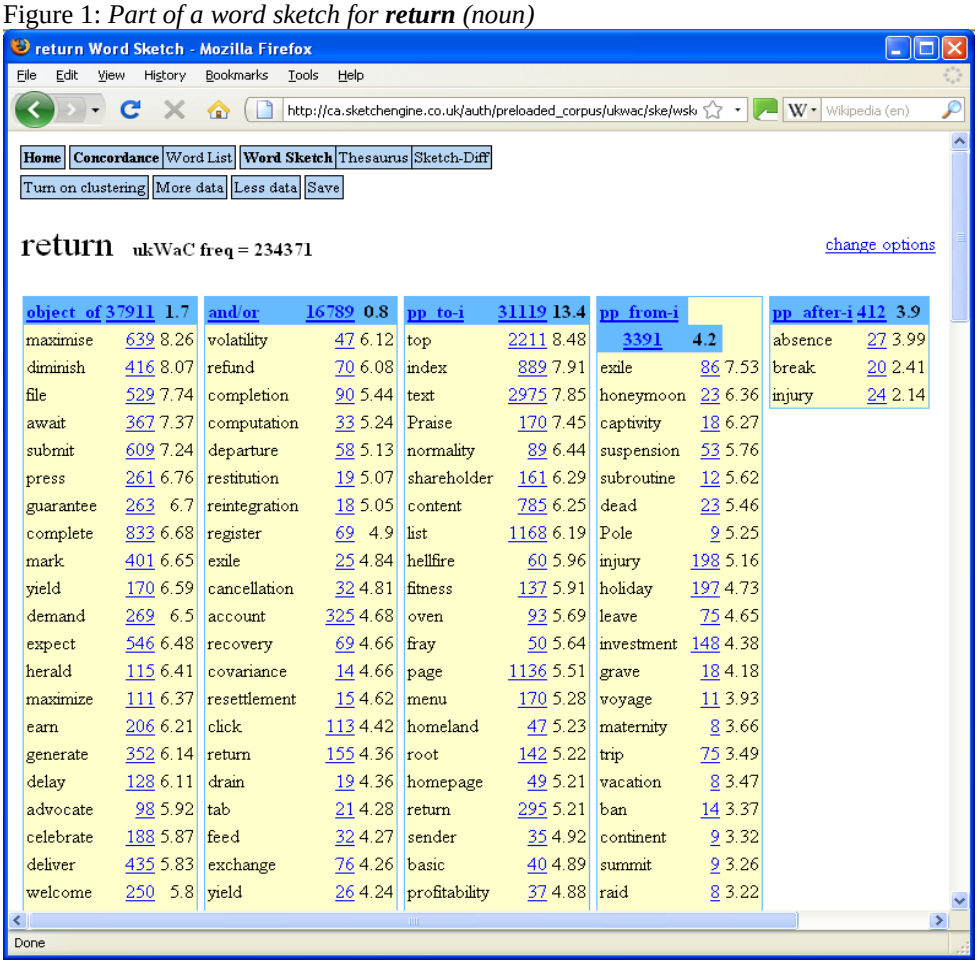
As in most areas of life, new ways of doing things typically evolve in response to known difficulties. What has tended to happen in the dictionary-development sphere is that we first identify a lexicographic problem, and then consider whether NLP techniques have anything to offer in the way of solutions. And when computational solutions are devised, we find – as often as not – that they have unforeseen consequences which go beyond the specific problem they were designed to address.

When planning a new dictionary, it is good to pay attention to what other dictionaries are doing, and to consider whether you can do the same things but do them better. But this is not enough. It is also important to look at emerging trends at the theoretical level and at their practical implications for language description. Collocation is a good example. The arrival of large corpora provided the empirical underpinning for a Firthian view of vocabulary, and – thanks to the work of John Sinclair and others – collocation became a core concept within the language-teaching community. Books such as Lewis (1993) and McCarthy & O'Dell (2005) helped to show the relevance of collocation at the classroom level, but in 1997 learner's dictionaries had not yet caught up: they showed an awareness of the concept, but their coverage of collocation was patchy and unsystematic. This represented an opportunity for MEDAL.

The first author described the problem to the second, who felt it should be possible to find all common collocations for all common words automatically, by using a shallow grammar to identify all verb-object pairs, subject-verb pairs, modifier-modifiee pairs and so on, and then to apply statistical filtering to give a fairly clean list, as proposed by Tapanainen & Järvinen (1998, and for the statistics, Church & Hanks 1989). The project would need a very large, part-of-speech-tagged corpus of general English: this had recently become available in the form of the British National Corpus. First experiments looked encouraging: the publisher contracted the researcher to proceed with the research,

and the first versions of word sketches were created. A word sketch is a one-page, corpus-based summary of a word's grammatical and collocational behaviour, as illustrated in Figure 1.

Figure 1: *Part of a word sketch for* ***return*** *(noun)*




| object_of 37911 1.7 | | and/or 16789 0.8 | | pp_to-i 31119 13.4 | | pp_from-i 3391 4.2 | | pp_after-i 412 3.9 | |
|---|---|---|---|---|---|---|---|---|---|
| maximise | 639 8.26 | volatility | 47 6.12 | top | 2211 8.48 | exile | 86 7.53 | absence | 27 3.99 |
| diminish | 416 8.07 | refund | 70 6.08 | index | 889 7.91 | honeymoon | 23 6.36 | break | 20 2.41 |
| file | 529 7.74 | completion | 90 5.44 | text | 2975 7.85 | captivity | 18 6.27 | injury | 24 2.14 |
| await | 367 7.37 | computation | 33 5.24 | Praise | 170 7.45 | suspension | 53 5.76 | | |
| submit | 609 7.24 | departure | 58 5.13 | normality | 89 6.44 | subroutine | 12 5.62 | | |
| press | 261 6.76 | restitution | 19 5.07 | shareholder | 161 6.29 | dead | 23 5.46 | | |
| guarantee | 263 6.7 | reintegration | 18 5.05 | content | 785 6.25 | Pole | 9 5.25 | | |
| complete | 833 6.68 | register | 69 4.9 | list | 1168 6.19 | injury | 198 5.16 | | |
| mark | 401 6.65 | exile | 25 4.84 | hellfire | 60 5.96 | holiday | 197 4.73 | | |
| yield | 170 6.59 | cancellation | 32 4.81 | fitness | 137 5.91 | leave | 75 4.65 | | |
| demand | 269 6.5 | account | 325 4.68 | oven | 93 5.69 | investment | 148 4.38 | | |
| expect | 546 6.48 | recovery | 69 4.66 | fray | 50 5.64 | grave | 18 4.18 | | |
| herald | 115 6.41 | covariance | 14 4.66 | page | 1136 5.51 | voyage | 11 3.93 | | |
| maximize | 111 6.37 | resettlement | 15 4.62 | menu | 170 5.28 | maternity | 8 3.66 | | |
| earn | 206 6.21 | click | 113 4.42 | homeland | 47 5.23 | trip | 75 3.49 | | |
| generate | 352 6.14 | return | 155 4.36 | root | 142 5.22 | vacation | 8 3.47 | | |
| delay | 128 6.11 | drain | 19 4.36 | homepage | 49 5.21 | ban | 14 3.37 | | |
| advocate | 98 5.92 | tab | 21 4.28 | return | 295 5.21 | continent | 9 3.32 | | |
| celebrate | 188 5.87 | feed | 32 4.27 | sender | 35 4.92 | summit | 9 3.26 | | |
| deliver | 435 5.83 | exchange | 76 4.26 | basic | 40 4.89 | raid | 8 3.22 | | |
| welcome | 250 5.8 | yield | 26 4.24 | profitability | 37 4.88 | | | | |

As the lexicographers became familiar with the software, it became apparent that word sketches did the job they were designed to do. Each headword's collocations could be listed exhaustively, to a far greater degree than was possible before. That was the immediate goal. But analysis of a word's sketch also tended to show, through its collocations, a wide range of the patterns of meaning and usage that it entered into. In most cases, each of a word's different meanings is associated with particular collocations, so the collocates listed in the word sketches provided valuable prompts in the key task of identifying and accounting for all the word's meanings in the entry. The word sketches functioned not only as a tool for finding collocations, but also as a useful guide to the

**Komentář [MU6]:** Or word sketch ? (AK – I prefer it as it is)

distinct senses of a word – the analytical core of the lexicographer's job (Kilgarriff & Rundell 2002).

Prior to the advent of word sketches, the primary means of analysis in corpus lexicography was the reading of concordances. Since the earliest days of the COBUILD project, the lexicographers scanned concordance lines – often in their thousands – to find all the patterns of meaning and use. The more lines were scanned, the more patterns would tend to be found (though with diminishing returns). This was good and objective, but also difficult and time-consuming. Dictionary publishers are always looking to save time, and hence budgets. Earlier efforts to offer computational support were based on finding frequently co-occurring words in a window surrounding the headword (Church & Hanks 1989). While these approaches had generated plenty of interest among university researchers, they were not taken up as routine processes by lexicographers: the ratio of noise to signal was high, the first impression of a collocation list was of a basket of earth with occasional glints of possible gems needing further exploration, and it took too long to use them for every word.

But early in the MEDAL project, it became clear that the word sketches were more like a contents page than a basket of earth. They provided a neat summary of most of what the lexicographer was likely to find by the traditional means of scanning concordances. There was not too much noise. Using them saved time. It was more efficient to start from the word sketch than from the concordance.

Thus the unexpected consequence was that the lexicographer's methodology changed, from one where the technology merely supported the corpus-analysis process, to one where it pro-actively identified what was likely to be interesting and directed the lexicographer's attention to it. And whereas, for a human, the bigger the corpus, the greater the problem of how to manage the data, for the computer, the bigger the corpus, the better the analyses: the more data there is, the better the prospects for finding all salient patterns and for distinguishing signal from noise. Though originally seen as a useful supplementary tool, the sketches provide a compact and revealing snapshot of a word's behaviour and uses and have, in most cases, become the preferred starting point in the process of analyzing complex headwords.

### *3.4 Word sketches and the Sketch Engine since 2004*

Since the first word sketches were used in the late 1990s in the development of the first edition of MEDAL, word sketches have been integrated into a general-purpose corpus query tool, the Sketch Engine (Kilgarriff et al. 2004) and have been developed for a dozen languages (the list is steadily growing). They are now in use for commercial and academic lexicography in the UK (where most of the main dictionary publishers use them), China, the Czech Republic, Germany, Greece, Japan, the Netherlands, Slovakia, Slovenia and the USA, and for language and linguistics teaching all round the world. Word sketches have been complemented by an automatic thesaurus (which identifies the words which are most similar, in terms of shared collocations, to a target word) and a range of other tools including 'sketch difference', for comparing and contrasting a word

with a near-synonym or antonym in terms of collocates shared and not shared. There are also options such as clustering a word's collocates or its thesaurus entries. The largest corpus for which word sketches have been created so far contains over five billion words (Pomikálek et al. 2009). In a quantitative evaluation, two thirds of the collocates in word sketches for five languages were found to be 'publishable quality': a lexicographer would want to include them in a published collocations dictionary for the language (Kilgarriff et al. 2010).

### *3.5 Word sketches and the Sketch Engine in the NEID project*

The New English-Irish Dictionary (NEID) is a project funded by Foras na Gaeilge, the statutory language board for Ireland, and planned by the Lexicography MasterClass.[10] It has provided a setting for a range of ambitious ideas about how we can efficiently create ever more detailed and accurate descriptions of the lexis of a language. The project makes a clear divide between the 'source-language analysis' phase of the project, and the translation and final-editing phases. A consequence is that the analysis phase is an analysis of English in which the target language (Irish) plays no part, and the resulting 'Database of ANalysed Texts of English' (DANTE) is a database with potential for a range of uses in lexicography and language technology. It could be used, for example, as a launchpad for bilingual dictionaries with a different target language, or as a resource for improving machine translation systems or text-remediation software. The Lexicography MasterClass undertook the analysis phase, with a large team of experienced lexicographers, over the period 2008-2010.[11]

The project has used the Sketch Engine with a corpus comprising UKWaC plus the English-language part of the New Corpus for Ireland (Kilgarriff et al. 2007). In the course of the project, three innovations were added to the standard word sketches.

#### *3.5.1 Customization of Sketch Grammar*

Any dictionary uses a particular grammatical scheme in its choice of the repertoire and meaning of the grammatical labels it attaches to words. The Sketch Engine also uses a grammatical scheme in its 'Sketch Grammar', which defines the grammatical relations according to which it will classify collocations in the word sketches: *object_of, and/or* etc. in Figure 1. The Sketch Grammar also gives names to the grammatical relations. This raises the prospect of mapping the grammatical scheme that is specified in a dictionary's Style Guide onto the scheme in the Sketch Grammar. In this way, there will be an exact match between the inventory of grammatical relations in the dictionary, and those presented to the lexicographer in the word sketch. A relation that is called NP_PP, for a verb such as *load* (*load the hay onto the cart*) in the lexical database will be called NP_PP, with exactly the same meaning, in the word sketch. Such an approach will simplify and rationalize the analysis process for the lexicographer: for the most part s/he will be copying a collocate of type X in the word sketch, into a collocate of type X (under the relevant sense of the headword) in the dictionary entry s/he is writing.

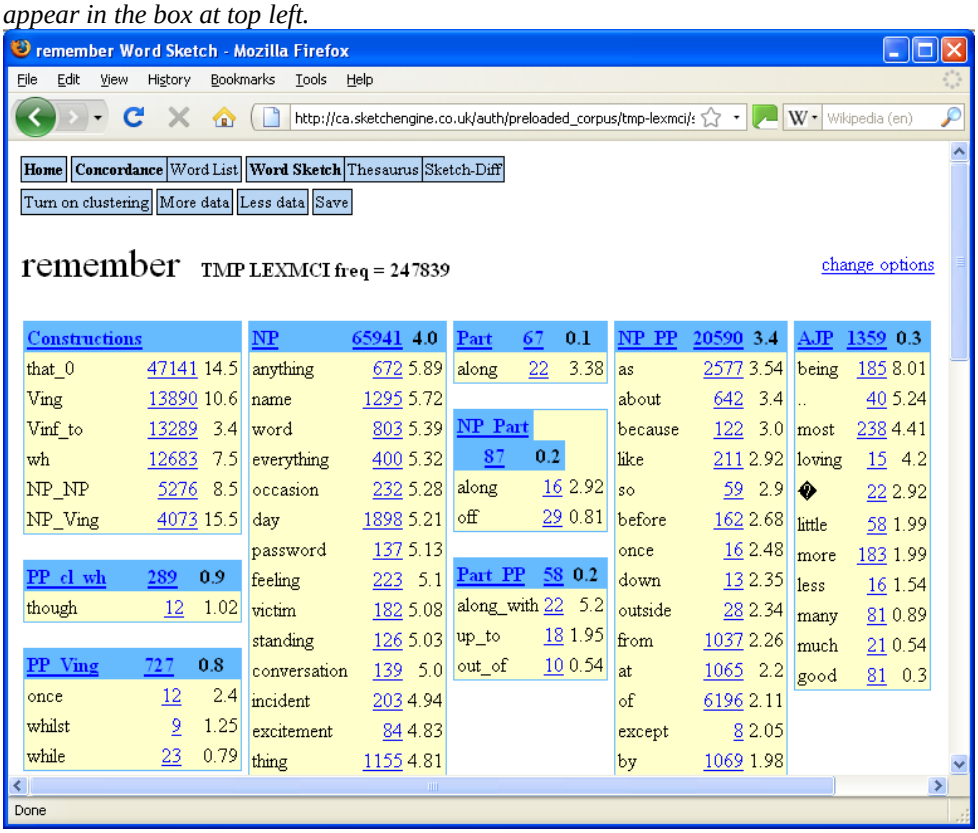
The NEID was the first project where the Sketch Grammar and Dictionary Grammar were fully harmonized: the Sketch Grammar was customized to express the same

grammatical constructions and collocation-types, with the same names, as the lexicographers would use in their analysis. Another Macmillan project (the *Macmillan Collocations Dictionary*; Rundell, ed., 2010) subsequently used the same approach.

*3.5.2 'Constructions list' as top-level summary of word sketch*

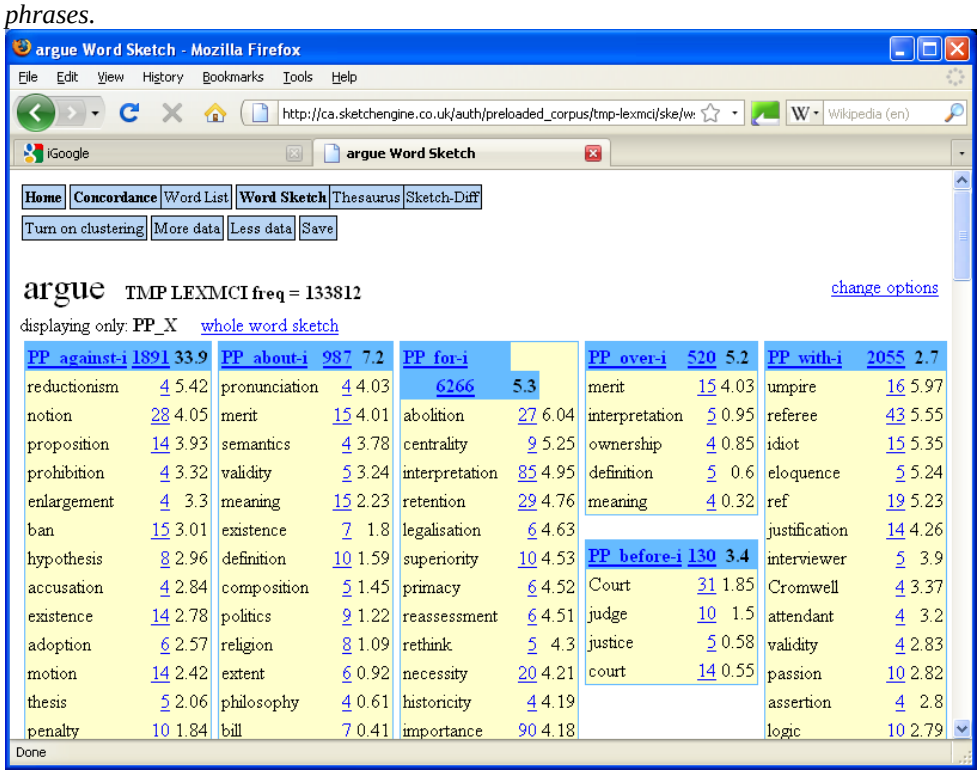
The dictionary grammar for the NEID project is quite complex and fine-grained. In the case of verbs, for example, any of 43 different structures may be recorded. Consequently we soon found that word sketches were often rather large and hard to navigate around. To address this, we introduced an 'index', which appears right at the top of the word sketch and summarizes its contents by listing the constructions that are most salient for that word (cf. Figure 2).

Figure 2: *Part of a word sketch for* ***remember*** *(verb). The verb's main syntactic patterns appear in the box at top left.*




In other cases, we found that there were a large number of constructions involving prepositions and particles, and that these could make the word sketch unwieldy. To address this, we collected all the preposition/particle relations on a separate web page, as in Figure 3.

Figure 3: *Word sketch for **argue**, showing part of the page devoted to prepositional phrases.*




### *3.5.3 'More data' and 'Less data' buttons*

The size of a word sketch is (inevitably) constrained by parameters which determine how many collocates and constructions are shown. The Sketch Engine has always allowed users to change the parameters, but most users are either unaware of the possibility or are not sure which parameters they should change or by how much. A simple but much-appreciated addition to the interface was 'More data' and 'Less data' buttons so the user can, at a single click, see less data (if they are feeling overwhelmed) or more data (if they have accounted for everything in the word sketch in front of them, but feel they have missed something or not said enough).

## *3.6 Labels*

Dictionaries use a range of labels (such as *usu pl.*, *informal*, *Biology*, *AmE*) to mark words according to their grammatical, register, domain, and regional-variety characteristics, whenever these deviate significantly from the (unmarked) norm. All of these are facts about a word's distribution, and all can, in principle, be gathered automatically from a corpus. In each of these four cases, computationalists are currently

able to propose some labels to the lexicographer, though there remains much work to be done.

In each case the methodology is to:

- specify a set of hypotheses
    - there will usually be one hypothesis per label, so grammatical hypotheses for the category 'verb' may include:
        - is it often/usually/always passive
        - is it often/usually/always progressive
        - is it often/usually/always in the imperative
- for each word
    - test all relevant hypotheses
    - for all hypotheses that are confirmed, add the information to the word sketch.

Where no hypotheses are confirmed – when, in other words, there is nothing interesting to say, which will be the usual case – nothing is added to the word sketch.

*3.6.1 Grammatical labels: usu. pl, usu. passive, etc.*

To determine whether a noun should be marked as 'usually plural', it is possible simply to count the number of times the lemma occurs in the plural, and the number of times it occurs overall, and divide the second number by the first to find the proportion. Similarly, to discover how often a verb is passivized, we can count how often it is a past participle preceded by a form of the verb *be* (with possible intervening adverbs) and determine what fraction of the verb's overall frequency the passive forms represent. Given a lemmatized, part-of-speech-tagged corpus, this is straightforward. A large number of grammatical hypotheses can be handled in this way.

The next question is: when is the information interesting enough to merit a label in a dictionary? Should we, for example, label all verbs which are over 50% passive as *often passive*?

To assess this question, we want to know what the implications would be: we do not want to bombard the dictionary user with too many labels (or the lexicographer with too many candidate-labels). What percentage of English verbs occur in the passive over half of the time? Is it 20%, or 50%, or 80%? This question is also not in principle hard to answer: for each verb, we work out its percentage passive, and sort according to the percentage. We can then give a figure which is, for lexicographic purposes, probably more informative than 'the percentage passive': the percentile. The percentile indicates whether a verb is in the top 1%, or 2%, or 5%, or 10% of verbs from the point of view of how passive they are. We can prepare lists as in Figure 4. This uses the methodology for finding the 'most passive' verbs (with frequency over 500) in the BNC. It shows that the most passive verb is *station*: people and things are often *stationed* in places, but there are far fewer cases where someone actively *stations* things. For *station*, 72.2% of its 557 occurrences are in the passive, and this puts it in the 0.2% 'most passive' verbs of English. At the other end of the table, *levy* is in the passive just over half the time, which puts it in the 1.9% most

passive verbs. The approach is similar to the collostructional analysis of Gries & Stefanowitsch (2004).

Figure 4: *The 'most passive' verbs in the BNC, for which a 'usually passive' label might be proposed.*

| Percentile | Ratio | Lemma | Frequency |
|---|---|---|---|
| 0.2 | 72.2 | station | 557 |
| 0.2 | 71.8 | base | 19201 |
| 0.3 | 71.1 | destine | 771 |
| 0.3 | 68.7 | doom | 520 |
| 0.4 | 66.3 | poise | 640 |
| 0.4 | 65.0 | situate | 2025 |
| 0.5 | 64.7 | schedule | 1602 |
| 0.5 | 64.1 | associate | 8094 |
| 0.6 | 63.2 | embed | 688 |
| 0.7 | 62.0 | entitle | 2669 |
| 0.8 | 59.8 | couple | 1421 |
| 0.9 | 58.1 | jail | 960 |
| 1.1 | 57.8 | deem | 1626 |
| 1.1 | 55.5 | confine | 2663 |
| 1.2 | 55.4 | arm | 1195 |
| 1.2 | 54.9 | design | 11662 |
| 1.3 | 53.9 | convict | 1298 |
| 1.5 | 53.1 | clothe | 749 |
| 1.5 | 52.8 | dedicate | 1291 |
| 1.5 | 52.4 | compose | 2391 |
| 1.6 | 51.5 | flank | 551 |
| 1.7 | 50.8 | gear | 733 |
| 1.9 | 50.1 | levy | 603 |

As can be seen from this sample, the information is lexicographically valid: all the verbs in the table would benefit from an *often passive* or *usually passive* label.

A table like this can be used by editorial policy-makers to determine a cut-off which is appropriate for a given project. For instance, what proportion of verbs should attract an *often passive* label? Perhaps the decision will be that users benefit most if the label is not overused, so just 4% of verbs would be thus labelled. The full version of the table in Figure 4 tells us what these verbs are. And now that we know precisely the hypothesis to use ("is the verb in the top 4% most-passive verbs?") and where the hypothesis is true, the label can be added into the word sketch. In this way, the element of chance – will the lexicographer notice whether a particular verb is typically passivized? – is eliminated, and the automation process not only reduces lexicographers' effort but at the same time ensures a more consistent account of word behaviour.

*3.6.2 Register Labels: formal, informal, etc.*

Any corpus is a collection of texts. Register is in the first instance a classification that applies to texts rather than words. A word is informal (or formal) if it shows a clear tendency to occur in informal (or formal) texts. To label words according to register, we need a corpus in which the constituent texts are themselves labelled for register in the document header. Note that at this stage, we are not considering aspects of register other than formality.

One way to come by such a corpus is to gather texts from sources known to be formal or informal. In a corpus such as the BNC, each document is supplied with various text type classifications, so we can, for example, infer from the fact that a document is everyday conversation, that it is informal, or from the fact that it is an academic journal article, that it is formal.

The approach has potential, but also drawbacks. In particular, it is not possible to apply it to any corpus which does not come with text-type information. Web corpora do not. An alternative is to build a classifier which infers formality level on the basis of the vocabulary and other features of the text. There are classifiers available for this task: see for example Heylighen & Dewaele (1999), and Santini et al. (2009). Following this route, we have recently labelled all documents in a five billion word web corpus according to formality, so we are now in a position to order words from most to least formal. The next tasks will be to assess the accuracy of the classification, and to consider – just as was done for passives – the percentage of the lexicon we want to label for register.

The reasoning may seem circular: we use formal (or informal) vocabulary to find formal (or informal) vocabulary. But it is a spiral rather than a circle: each cycle has more information at its disposal than the previous one. We use our knowledge of the words that are formal or informal to identify documents that are formal or informal. That then gives us a richer dataset for identifying further words, phrases and constructions which tend to be formal or informal, and allows us to quantify the tendencies.

*3.6.3 Domain Labels: Geol., Astron., etc*

The issues are, in principle, the same as for register. The practical difference is that there are far more domains (and domain labels): even MEDAL, a general-purpose learner's dictionary, has 18 of these, while the NEID database has over 150 domain labels. Collecting large corpora for each of these domains is a significant challenge.

It is tempting to gather a large quantity of, for example, geological texts from a particular source, perhaps an online geology journal. But rather than being a 'general geology' corpus, that subcorpus will be an 'academic-geology-prose corpus', and the words which are particularly common in the subcorpus will include vocabulary typical of academic discourse in general as well as of the domain of geology. Ideally, each subcorpus will have the same proportions of different text-types as the whole corpus. None of this is technically or practically impossible, but the larger the number of subcorpora, the harder it is to achieve.

In current work, we are focusing on just three subcorpora: legal, medical and business, to see if we can effectively propose labels for them.

Once we have the corpora and counts for each word in each subcorpus, we need to use statistical measures for deciding which words are most distinctive of the subcorpus: which words are its 'keywords', the words for which there is the strongest case for labelling. The maths we use is based on a simple ratio between relative frequencies, as implemented in the Sketch Engine and presented in Kilgarriff (2009).

*3.6.4 Region Labels: AmE, AustrE, etc*

The issues concerning region labels are the same as for domains but in some ways a little simpler. The taxonomy of regions, at least from the point of view of labelling items used in different parts of the English-speaking world, is relatively limited, and a good deal less open-ended than the taxonomy of domains. In MEDAL, for example, it comprises just 12 varieties or dialects, including American, Australian, Irish, and South African English.

*3.7 Examples*

Most dictionaries include example sentences. They are especially important in pedagogical dictionaries, where a carefully-selected set of examples can clarify meaning, illustrate a word's contextual and combinatorial behaviour, and serve as models for language production. The benefits for users are clear, and the shift from paper to electronic media means that we can now offer users far more examples. But this comes at a cost. Finding good examples in a mass of corpus data is labour-intensive. For all sorts of reasons, a majority of corpus sentences will not be suitable as they stand, so the lexicographer must either search out the best ones or modify corpus sentences which are promising but in some way flawed.

*3.7.1 GDEX*

In 2007, the requirement arose – in a project for Macmillan – for the addition of new examples for around 8,000 collocations. The options were to ask lexicographers to select and edit these in the 'traditional' way, or to see whether the example-finding process could be automated. Budgetary considerations favoured the latter approach, and subsequent discussions led to the GDEX ('good dictionary examples') algorithm, which is described in Kilgarriff et al. (2008).

Essentially, the software applies a number of filters designed to identify those sentences in a corpus which most successfully fulfil our criteria for being a 'good' example. A wide range of heuristics is used, including criteria like sentence length, the presence (or absence) of rare words or proper names, and the number of pronouns in the sentence. The system worked successfully on its first outing – not in the sense that every example it 'promoted' was immediately usable, but in the sense that it significantly streamlined the lexicographer's task. GDEX continues to be refined, as more selection criteria are added and the weightings of the different filters adjusted. For the DANTE database, which includes several hundred thousand examples, GDEX sorts the sentences for any of the combinations shown in the word sketches, in such a way that the ones which GDEX

thinks are 'best' are shown first. The lexicographer can scan a short list until they find a suitable example for whatever feature is being illustrated, and GDEX means they are likely to find what they are looking for in the top five examples, rather than, on average, within the top 20 to 30.

*3.7.2 One-click copying*

DANTE is an example-rich database in which almost all word senses, constructions, and multiword expressions are illustrated with at least one example. All examples are from the corpus and are unedited (DANTE is a lexical database rather than a finished dictionary). Lexicographers are thus required to copy many example sentences from the corpus system into the dictionary editing system. We use standard copy-and-paste but in the past this has often been fiddly, with one click to see the whole sentence, then manoeuvring the mouse to mark it all. So we have added a button for 'one-click copying': now, a single click on an icon at the right of any concordance line copies not the visible concordance line, but the complete *sentence* (with headword highlighted) and puts it on the clipboard ready for pasting into the dictionary.

*3.8 Tickbox lexicography (TBL)*

One-click copying is a good example of a simple software tweak that streamlines a routine lexicographic task. This may look trivial, but in the course of a project such as DANTE, the lexicographic team will be selecting and copying several hundred thousand example sentences, so the time-savings this yields are significant.

Another development – currently in use on two lexicographic projects – takes this process a step further, allowing lexicographers to select collocations for an entry, then select corpus examples for each collocation, simply by ticking boxes (thus eliminating the need to retype or cut-and-paste). We call this 'tickbox lexicography' (TBL), and in this process, the lexicographer works with a modified version of the word sketches, where each collocate listed under the various grammatical relations ('gramrels') has a tickbox beside it. Then, for each word sense and each gramrel, the lexicographer:

- ticks the collocations s/he wants in the dictionary or database
- clicks a 'Next' button
- is then presented with a choice of six corpus examples for every collocation, each with a tickbox beside it (six is the default, and assumes that – thanks to GDEX – a suitable example will appear in this small set; but the defaults can of course be changed)
- ticks the desired examples, then clicks a 'Copy to clipboard' button.

The system then builds an XML structure according to the DTD (Document Type Definition) of the target dictionary (each target dictionary has its own TBL application). The lexicographer can then paste this complex structure, in a single move, directly into the appropriate fields in the dictionary writing system. In this way, TBL models and streamlines the process of getting a corpus analysis out of the corpus system and into the dictionary writing system, as the first stage in the compilation of a dictionary. Here again, the incremental efficiency gains are substantial. The TBL process is especially well-

adapted to the emerging situation where online dictionaries give their users access to multiple examples of a given linguistic feature (such as a collocation or syntax pattern): with TBL, large numbers of relevant corpus examples can be selected and copied into the database with minimum effort.

## 4. Conclusions

If we look back at the list of lexicographic tasks (Section 3, above), we find that the following have been – or soon will be – automated to a significant degree:

- corpus creation
- headword list building
- identification of key linguistic features or preferences (syntactic, collocational, colligational, and text-type-related)
- example selection.

Further improvements are possible for each of these technologies (notably the GDEX algorithm and the text-type classifiers), and many of these are already in development. An especially interesting approach we are now looking at is one that takes the whole automation process a step further. In this model, we envisage a change from the current situation, where the corpus software (some version of the word sketches) presents data to the lexicographer in (as we have seen) intelligently pre-digested form, to a new paradigm where the software selects what it believes to be relevant data and actually populates the appropriate fields in the dictionary database. In this way of working, the lexicographer's task changes from selecting and copying data from the software, to validating – in the dictionary writing system – the choices made by the computer. Having deleted or adjusted anything unwanted, the lexicographer then tidies up and completes the entry. The principle here is that, assuming the software can be trained to make the 'right' decisions in a majority of cases, it is more efficient to edit out the computer's errors than to go through the whole data-selection process from the beginning. If this approach can be made to work effectively, we are likely to see a further change in lexicographers' working practices – and a further shift towards full automation.

Automated lexicography is still some way off. In particular, we have not yet reached the point where definition writing and (hardest of all) word sense disambiguation (WSD) are carried out by machines. In both cases, however, it may be possible to solve the problem by redefining the goal. If, for example, we think less in terms of discreet, numbered 'dictionary senses', and more of the contribution that a word makes to the meaning of a given communicative event, then the task starts to look less intractable. It has become increasingly clear that the meaning of a word in a particular context is closely associated with the specific patterning in which it appears – where 'patterning' encompasses features such as syntax, collocation, and domain information. A good deal of research is going on in this area, notably Patrick Hanks' work on 'Corpus Pattern Analysis' (e.g. Hanks 2004), and it is self-evident that computers can identify and count clusters of patterns more readily than they can count something as unstable as 'senses'. This offers one way forward. Equally, definitions could become less important if the user who encounters an unknown word could immediately access half a dozen very similar corpus

examples (filtered by GDEX or the like), and then draw his or her own conclusions. Whether this could be a viable alternative to the traditional definition – especially when the user is a learner – remains to be seen.

We have described a long-running collaboration between a lexicographer and a computational linguist, and its outcomes in terms of the way that dictionary text is compiled in the early 21st century. There is plenty more to be done, but it should be clear from this brief survey that the interaction between lexicography and language engineering has already been fruitful and promises to deliver even greater benefits in the future.

## Notes

[1] We are aware that our detailed knowledge relates mainly to developments in English-language lexicography. We apologise in advance for our Anglocentrism and any exaggerated claims it has led to.
[2] We should perhaps add this rider: "at least for the most widely-used languages, for which many billions of words of text are now available".
[3] "Every time COBUILD doubles its corpus, we want to double it again" (Clear 1996: 266).
[4] Hence, for example, there are now substantial corpora for 'smaller' languages such as Irish or the Bantu languages of southern Africa: Kilgarriff, et al. (2006), de Schryver & Prinsloo (2000).
[5] See for example Keller & Lapata (2003), Fletcher (2004). For general background to web corpora, see Kilgarriff & Grefenstette (2003), Atkins & Rundell (2008: 78-80), Baroni et al. (2009).
[6] In the BNC *mucosa* is marginally more frequent than *spontaneous* and *enjoyment*, though of course it appears in far fewer corpus documents.
[7] As is now generally recognised, the notion of 'representativeness' is problematical with regard to general-purpose corpora like BNC and UKWaC, and there is no 'scientific' way of achieving it: see e.g. Atkins & Rundell (2008: 66).
[8] The issue came to our attention when an early version of the BNC frequency list gave undue prominence to verbal *car*.
[9] Here we exclude inflectional morphemes, addressed under lemmatization above: in English a distinction between inflectional and derivational morphology is easily made.
[10] http://www.lexmasterclass.com.
[11] For an account see Atkins et al. (2010).

## References


Atkins, S., Kilgarriff, A., & Rundell, M. 2010. The Database of Analysed Texts of English (DANTE). *Proceedings of 14th EURALEX International Congress*, A. Dykstra & T. Schoonheim (eds). Leeuwarden, The Netherlands.

Atkins, S. & Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baroni, M. & Bernardini, S. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*: 1313-1316. Lisbon.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal* 43(3): 209-226.

Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. 2006. WebBootCaT: A Web Tool for Instant Corpora. In *Proceedings of 12th EURALEX International Congress*, E. Corino, C. Marello, C. Onesti (eds), 123-131. Alessandria: Edizioni Dell'Orso.

Church, K. & Hanks, P. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16:22–29.

Clear, J. 1988. The Monitor Corpus. In *ZüriLEX '86 Proceedings*, M. Snell-Hornby (ed.), 383-389. Tübingen: Francke Verlag.

Clear, J. 1996. Technical Implications of Multilingual Corpus Lexicography. *International Journal of Lexicography* 9(3): 265-273.

de Schryver, G-M & Prinsloo, D. J. 2000. The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies* 18(1-4): 89-106.

Fairon, C., Macé, K., & Naets, H. 2008. GlossaNet2: a linguistic search engine for RSS-based corpora. *Proceedings, Web As Corpus Workshop (WAC4),* S. Evert, A. Kilgarriff & S. Sharoff (eds), 34-39. Marrakesh.

Fletcher, W. H. 2004. Making the Web More Useful as a Source for Linguistic Corpora. In *Applied Corpus Linguistics*: *A Multidimensional Perspective*, U. Connor & T. Upton (eds), 191-205. Amsterdam: Rodopi.

Grefenstette, G. 1998. The Future of Linguistics and Lexicographers: Will there be Lexicographers in the Year 3000? In *Actes EURALEX 1998*, T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin & S. Theissen (eds), 25-42. Liège: Université de Liège.

Gries, S. Th. & Stefanowitsch, A. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97-129.

Hanks, P. W. 2004. Corpus Pattern Analysis. In *Proceedings of the Eleventh Euralex Congress*, G. Williams & S. Vessier (eds), 87-98. Lorient, France: UBS.

Heylighen F. & Dewaele, J.-M. 1999. Formality of Language: Definition, measurement and behavioural determinants. Internal Report, Free University Brussels, http://pespmc1.vub.ac.be/Papers/Formality.pdf

Janicivic, T. & Walker, D. 1997. NeoloSearch: Automatic Detection of Neologisms in French Internet Documents. *Proceedings of ACH/ALLC'97*: 93-94. Queen's University, Ontario, Canada.

Keller, F. & Lapata, M. 2003. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics* 29(3): 459-484.

Kilgarriff, A. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography* 10(2): 135-155.

Kilgarriff, A. 2006. Collocationality (and how to Measure it). In *Proceedings of 12th EURALEX International Congress*, E. Corino, C. Marello & C. Onesti (eds), 997-1004. Alessandria: Edizioni Dell'Orso.

Kilgarriff, A. 2009. Simple maths for keywords. *Proceedings, Corpus Linguistics.* M. Mahlberg, V. González-Díaz & C. Smith (eds). Liverpool; online at http://ucrel.lancs.ac.uk/publications/cl2009/.

Kilgarriff, A. 2010. Comparable corpora within and across languages, word frequency lists and the Kelly project. *Proceedings, 3rd Workshop on Building and Using Comparable Corpora*. R. Rapp, P. Zweigenbaum & S. Sharoff (eds). LREC, Malta.

Kilgarriff, A. & Grefenstette, G. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3): 333-348.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the XIII Euralex Congress*, E. Bernal & J. DeCesaris (eds), 425-431. Barcelona: Universitat Pompeu Fabra.

Kilgarriff, A., Kovář, V. Krek, S. Srdanović, I., & Tiberius, C. 2010. A quantitative evaluation of word sketches. *Proceedings of 14th EURALEX International Congress*, A. Dykstra & T. Schoonheim (eds). Leeuwarden, The Netherlands.

Kilgarriff, A. & Rundell, M. 2002. Lexical Profiling Software and its Lexicographic Applications: A Case Study. In *Proceedings of the Tenth Euralex Congress*, A. Braasch & C. Povlsen (eds), 807-818. Copenhagen: University of Copenhagen.

Kilgarriff, A., Rundell, M., & Uí Dhonnchadha, E. 2006. Efficient corpus development for lexicography: Building the New Corpus for Ireland. *Language Resources and Evaluation Journal* 40(2): 127-152.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. 2004. The Sketch Engine In *Proceedings of the Eleventh Euralex Congress*, G. Williams & S. Vessier (eds), 105-116. Lorient, France: UBS.

Krishnamurthy, R. 1987. The Process of Compilation. In Sinclair J. M. (ed.). *Looking Up*: *An Account of the COBUILD Project in Lexical Computing*. London: Collins. Pp 62-85.

Kučera, H. & Francis, W. N. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Lewis, M. 1993. *The Lexical Approach*. Hove, UK: Language Teaching Publications.

McCarthy, M. & O'Dell, F. 2005. *English Collocations in Use*. Cambridge: Cambridge University Press.

Murray, K. E .M. 1979. *Caught in the Web of Words*: *James A.H. Murray and the Oxford English Dictionary*. Oxford: Oxford University Press.

Murray, J., Bradley, H., Craigie, W. & Onions, C. T. 1928. *Oxford English Dictionary.* Oxford: Oxford University Press.

O'Donovan, R. & O'Neill, M. 2008. A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In *Proceedings of the XIII Euralex Congress*, E. Bernal & J. DeCesaris (eds), 571-579. Barcelona: Universitat Pompeu Fabra.

Pomikálek, J., Rychlý, P. & Kilgarriff, A. 2009. Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics. Special Issue of *Research in Computing Science* 41: Mexico City.

Procter, P. (ed.). 1978. *Longman Dictionary of Contemporary English*. Harlow: Longman.

Renouf, A. 1987. 'Corpus Development'. In Sinclair J. M. (ed.). *Looking Up*: *An Account of the COBUILD Project in Lexical Computing*. London: Collins:10-40.

Rundell, M. (ed.). 2001. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Education.

Rundell, M. (ed.). 2010. *Macmillan Collocations Dictionary*. Oxford: Macmillan Education.

Santini M., Rehm, G., Sharoff, S., & Mehler, A. (eds). 2009. Introduction, *Journal for Language Technology and Computational Linguistics*, Special Issue on Automatic Genre Identification: Issues and Prospects. 24(1):129-145.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In Baroni, M. & Bernardini, S. (eds). *Wacky! Working Papers on Web as Corpus.* Bologna: Gedit..

Stein, J. & Urdang, L. (eds). 1966. *Random House Dictionary of the English Language*. New York: Random House Inc.

Tapanainen, P. & Järvinen, T. 1998. Dependency Concordances. International *Journal of Lexicography* 11(3):187-203.