Lexical Profiling Software and its Lexicographic Applications – a Case Study

Adam Kilgarriff and Michael Rundell
Information Technology Research Institute, University of Brighton and the Lexicography MasterClass
adam.kilgarriff@itri.brighton.ac.uk, michael.rundell@itri.brighton.ac.uk

Abstract
The latest generation of lexical profiling software (which developed out of the probability measures originally proposed by Church and Hanks) has recently been used as a central source of linguistic data for a new, written-from-scratch pedagogical dictionary. The "Word Sketch" software uses parsed corpus data to identify salient collocates – in separate lists – for the whole range of grammatical relations in which a given word participates. It also links these collocate lists to corpus examples instantiating each combination so identified. Lexicographers found that the Word Sketches not only streamlined the process of searching for significant word combinations, but often provided a more revealing, and more efficient, way of uncovering the key features of a word's behaviour than the (now traditional) method of scanning concordances.

1. Introduction

The debate, in corpus lexicography, has moved on from the issue of whether to use a corpus at all (the 1980s), through questions of corpus size and corpus "representativeness" (the 1990s), to the issue of how to extract maximum value from corpus resources. As corpora grow, so the number of corpus lines for a word grows, and the lexicographer needs a solution to the problem of information overload (section 2). Statistical summaries offer a way forward, though until recently they have usually been limited in their usefulness, in part because they have usually been "grammatically blind" (section 3). In the work presented in this paper, we present the Word Sketch, a new response to these challenges (sections 4, 5) and describe how Word Sketches have been used in a large lexicographic project and what lessons have been learnt from that (section 6).

2. Information overload

Investigating lexical behaviour in general, and combinatorial behaviour in particular, requires very large volumes of text. Available technology can now supply these needs, for English at least, without the major – even heroic – efforts that characterized the early days of lexicographic corpus building. But this in turn brings "information overload" problems for lexicographers. Scanning concordance lines, the "traditional" approach to analyzing corpus data, begins to make unreasonable demands on human memory once the number of instances we need to look at goes above about 300. Yet a 200-million-word corpus (not large by today's standards) will supply two or three times that number of concordance lines even for words of very modest frequency (such as **abrupt**, **accentuate**, and **accompaniment**), while for anything more central (words like **abandon**, **absorb**, or **absolute**) we could be looking at several thousand lines. The problem here is not simply that this is very time-consuming (and therefore unlikely to be feasible within the normal constraints of commercial publishing), but that human editors cannot process such high volumes of data with any degree of reliability.

In a single generation, we have gone from famine to feast. Lexicographers on the first COBUILD project (in the early 1980s) worked with a corpus of not much over 7 million words – and often found themselves wanting more. Twenty years on, we are almost drowning in data, and a 100-million-word corpus would now be seen, by most English dictionary publishers, as no more than "entry-level". The major requirement therefore is for software tools that can fully exploit the benefits of very large corpora, while preserving lexicographers from an excess of information. The need, broadly, is for some form of automated summarizing utility that will present dictionary-writers with a pre-digested outline of the most important and relevant facts about a word. The precise form that such a tool might take is not yet clear. The simplest procedure, of course, is to take a sample of the available data, and most corpus-querying tools allow users to request (say) 500 randomly chosen concordance lines when many thousands are available. But this is not a real solution: arguably, taking a sample negates the value of having a large corpus, and for the lexicographer there is always the concern that vital data may have slipped through the "sieve" when the sampling was done.

3. Statistical summaries

As corpora grew ever larger, Church and Hanks [1989] opened up a promising new avenue with their proposal for the use of statistical measures of co-occurrence as a way of automatically identifying significant collocations.

The method described by Church and Hanks is essentially as follows (the term "nodeword" here refers to the word whose combinatorial behaviour is being investigated):

- for each corpus instance of the nodeword, find all words occurring within *k* words of it; keep a tally for each co-occurring word
- for each such co-occurring word, compute a statistic to measure how noteworthy the relation between it and the nodeword is
- sort words according to the statistic, showing lexicographers only the items with the highest scores

Statistics vary according to how they assess and measure noteworthiness. This is done by finding how *improbable* the collocation is, given the probabilities of each of its component words. Probabilities are estimated on the basis of corpus frequencies. The challenge for the mathematician is to accurately estimate and compare the probabilities, given the frequency data.

Church and Hanks presented the Mutual Information (MI) statistic,

$MI(x ,y) = \log_2(P(x,y) / (P(x).P(y)))$

Here, *x, y* are the words forming what might be a collocation; *P(x, y*) is the probability of the two words occurring together, and *P(x)* and *P(y)* are the probabilities of each word occurring irrespective of the collocate. Probabilities are estimated from corpus frequencies simply by dividing the frequency by the size of the corpus, so a word occurring 1000 times in a million-word-corpus has a probability of 1000/1,000,000 = 1/1000.

If we assume that there is no particular link between the two words (the so-called "null hypothesis"), then we can predict the frequency with which they will *co-occur* in the corpus from the frequency with which each occurs independently. For example, if each word occurs, on average, once per thousand words, we would expect the first to come immediately before the second just once per million words: according to the definition of statistical independence, if two events are independent, then the probability of them occurring together is the product of their probabilities. Conversely, if the relation between the words is noteworthy, they will appear together far more often than this. MI measures noteworthiness by calculating how many times more than the expected value (here, one per million) the words co-occur.

T-score, introduced to the corpus linguistics community by Gale et al. [1991], works on a similar basis but adds the information that larger counts support more accurate estimates of probabilities than small counts. When used to measure the noteworthiness of one word in relation to one other[1], it measures the number of standard deviations between observed and expected frequencies of a collocation, given the independent frequencies of each collocate. The log-likelihood statistic, introduced by Dunning [1992], is similar to MI but makes allowance for the unreliability of estimates of noteworthiness based on very low counts. (MI tends to overstate the noteworthiness of collocations where at least one of the co-occurring words is itself somewhat rare.) In a similar vein, Pedersen [1996] shows how probabilities can be calculated exactly even where counts are low.

For lexicographers, probability measures like these appeared to offer a solution to the "information overload" problem: concordances would now be complemented by a statistical summary that revealed, at a glance, the salient facts about a word's combinatory preferences. Consequently, Church and Hanks' paper caused considerable excitement in the lexicographic community, and statistical measures of this type did indeed quickly become a standard feature of many of the corpus-querying tools used by dictionary writers: programs such as Corpus Bench, WordSmith Tools, and QWICK all incorporate various forms of statistically-based collocation-listing tool.

Yet in practice such tools have not, on the whole, become a standard part of the lexicographic process[2], and one is bound to wonder why this should be. Lexicographers never have enough time, so will only consult those sources that deliver significantly "better" data. This has been manifest where corpora have first become available. Scanning concordances is substantially replacing more traditional methods of viewing evidence. The new approach requires more time, but the payoff in terms of improved linguistic information is high. But statistical summaries, despite a high level of initial interest among the dictionary community, have had a far more limited impact.

The essential problem with these collocate lists is that they are "noisy". That is – while they are certainly suggestive and can sometimes nudge editors in useful directions – they require too much interpretation to be genuinely useful as a standard lexicographic tool. Too much of the information they present is either irrelevant or misleading, so a good deal of human intervention is required in order to extract data of real value.

To illustrate some of the issues, consider the output of a search made by the COBUILD Online Collocation Sampler. The COBUILD website offers a collocation-listing service based on a 56-million-word subset of the Bank of English. Lists of statistically-significant collocates can be requested for a given nodeword, and users can choose either the MI measure or the T-score. The following table shows the ten most "significant" collocates of the word **conversation** using each of these measures:

MI Score	*T-score*
overhearing	with
phatic	a
overhear	had
eavesdrop	in
snatches	telephone
stilted	between
transcripts	our
overheard	about
topic	into
peppered	phone

Table 1: Comparing MI and T-scores for **conversation**

The differences between the two lists are striking. As noted above, MI gives undue weight to collocates which are themselves very infrequent words: the high end of MI lists therefore tend to be populated with quite unusual items. The word **phatic** (which appears just 8 times in the BNC's 100 million words) is the most egregious example here, but **stilted** and **peppered** are also quite surprising members of a list of the top ten collocates of **conversation**, and few lexicographers would argue for taking account of *any* of these words in an entry for **conversation**. The T-score measure, conversely, makes adjustments that take account of the size of the joint frequency figure: this smooths out many of the problems associated with MI, but has its own disadvantages in that it gives high significance scores to extremely common words. It may be useful to be reminded of the prepositions that usually follow **conversation**, but one does not need sophisticated software tools to be told that the indefinite article co-occurs frequently with this word. In practice, lexicographically-interesting information tends to be found in the middle reaches of most T-score lists rather than at the very top, so here again, extracting useful information requires a certain amount of persistence.

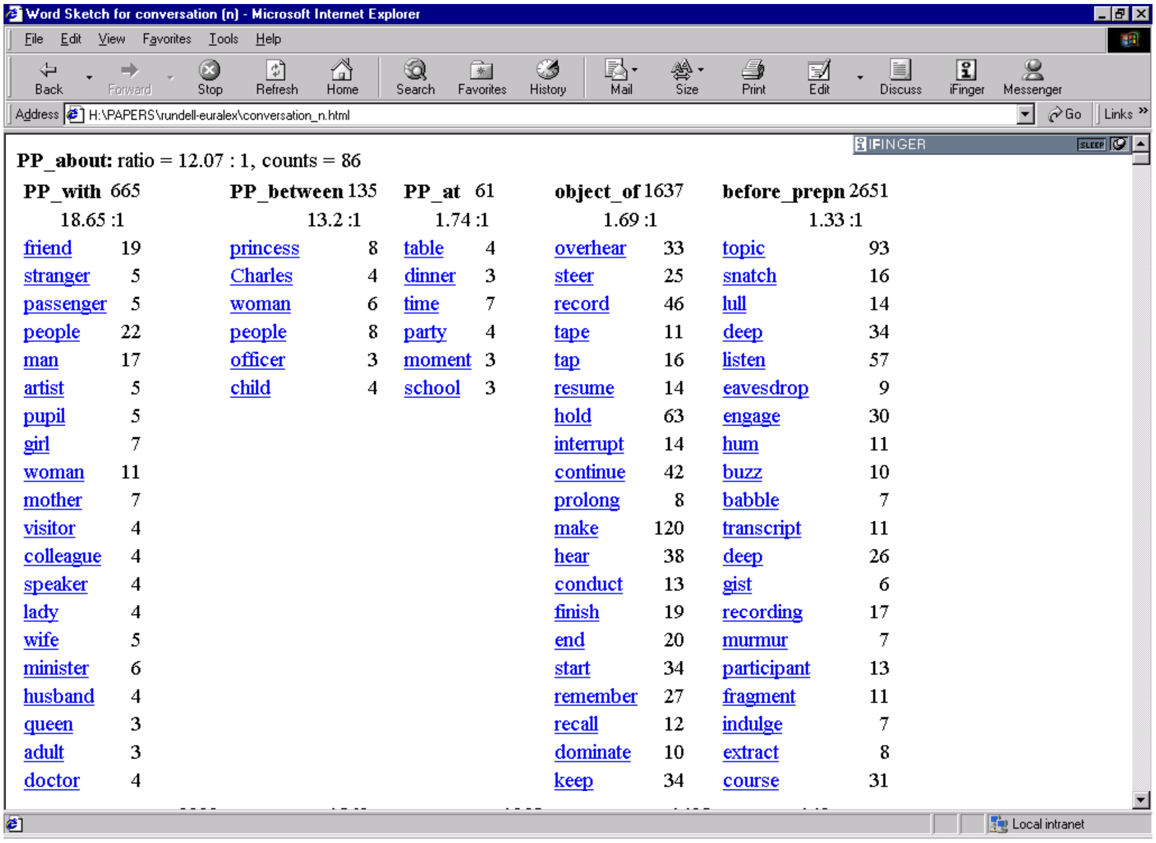
Two further problems relate to lemmatization and window size. Regarding lemmatization, the value of the data is limited by the fact that the software simply identifies individual *word forms* rather than whole *lemmas*. For example, the MI list above shows three parts of the lemma **overhear**, but only one of the lemma **eavesdrop**. But what lexicographers need to be able to do is compare the complete co-occurrence frequencies of these two verbs – any manual attempt to work this out would take too long and be of doubtful reliability. Or again, the word **snatches** could be either a third-person-singular present tense verb or a plural noun – a distinction that a genuinely useful system should be able to make.

Regarding window-size, software such as CorpusBench allows the lexicographer to choose the window in which collocates are to be sought, so lists can be generated for "immediately preceding word" or "any of the three following words" or "all words within five words of the nodeword, preceding or following". Different windows show different kinds of information: small windows tend to call up grammatical collocates, larger ones, lexical ones. This leaves the lexicographer with far too many: how many different collocates lists should be specified, called up, and examined for a given nodeword? The question adds extra work for the lexicographer.

In their basic form, then, lists of this type are usefully suggestive but contain too much noise, require too much interpretation, and are too arbitrary in how they are specified, to be an indispensable lexicographic tool.

4. Word Sketches

The significance of Church and Hanks' paper and ensuing work was that it pointed the way to a new generation of lexical profiling software of a more sophisticated type, which would address some of the shortcomings of their original methods. In Stuttgart, Heid and colleagues have been developing such software using German corpora [Heid et al., 2000]. In Brighton, we have developed "Word Sketches" for English. The Word Sketches aim to improve on existing collocate lists by using POS-tagged and (partially) parsed corpus data to identify the salient collocates for a range of distinct grammatical relations. Thus, in place of the grammatically blind lists shown above, where nouns, verbs, adjectives, and prepositions are all lumped together, the Word Sketches provide *separate* collocate lists for different grammatical patterns. The Word Sketch for conversation, for example, lists – among many other combinations – verbs used when conversation is in the object position (such as overhear, steer, resume, and interrupt), verbs used when conversation is the subject (such as drift, cease, veer, and wander), and nouns appearing in the pattern NOUN + PREP/of + conversation (such as topic, snatch, hum, and buzz). For every collocate listed, there is a link to a set of example sentences from the corpus that show the pattern in use.


Word Sketch for conversation (n) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Size Print Edit Discuss iFinger Messenger
Address H:\PAPERS\rundell-euralex\conversation_n.html
Go Links
IFINGER
Local intranet

PP_about: ratio = 12.07 : 1, counts = 86

PP_with	665	**PP_between**	135	**PP_at**	61	**object_of**	1637	**before_prepn**	2651
18.65 :1		13.2 :1		1.74 :1		1.69 :1		1.33 :1	
friend	19	princess	8	table	4	overhear	33	topic	93
stranger	5	Charles	4	dinner	3	steer	25	snatch	16
passenger	5	woman	6	time	7	record	46	lull	14
people	22	people	8	party	4	tape	11	deep	34
man	17	officer	3	moment	3	tap	16	listen	57
artist	5	child	4	school	3	resume	14	eavesdrop	9
pupil	5					hold	63	engage	30
girl	7					interrupt	14	hum	11
woman	11					continue	42	buzz	10
mother	7					prolong	8	babble	7
visitor	4					make	120	transcript	11
colleague	4					hear	38	deep	26
speaker	4					conduct	13	gist	6
lady	4					finish	19	recording	17
wife	5					end	20	murmur	7
minister	6					start	34	participant	13
husband	4					remember	27	fragment	11
queen	3					recall	12	indulge	7
adult	3					dominate	10	extract	8
doctor	4					keep	34	course	31

Figure 1: Extract from Word Sketch for **conversation**

The Word Sketches developed to date have used the British National Corpus (http://info.ox.ac.uk/bnc) as the source of corpus data.

4.1 NLP technologies

Word Sketches were developed as part of a project aiming to bring together corpus lexicography and NLP (Natural Language Processing, also known as computational linguistics, language engineering, human language technologies or HLT). The most salient technologies are tokenization, lemmatization, part-of-speech tagging and parsing[3].

Tokenization is the process of identifying the words, by identifying characters and character sequences that occur within words and ones that occur between words. This is largely straightforward for English and other European languages, though hyphens and compounds present challenges.

Lemmatization is the process of identifying, e.g., **abduct** *(v)* as the lemma for the individual graphic forms **abduct, abducted, abducting, abducts**. Part-of-speech tagging is the process of identifying, for an ambiguous item such as **lapses**, whether it is, in a particular context, a plural noun or a third-person-singular, present-tense verb. Parsing is, in this context, the process of identifying grammatical relations between lexical items, to find that, e.g., in "*Dog bites man*" **man** is the object of **bite**, whereas in "*Man bites dog*", it is the subject. In general, parsing concerns the identification of relations between sentence-parts, but our focus is narrower: we are typically concerned with relations applying to heads of noun and verb phrases, rather than to the noun and verb phrases in their entirety. Thus, in "*The conversation had lapsed",* the relation we wish to note is between the lemmas **conversation** and **lapse***,* not between the noun phrase *the conversation* and the verb phrase *had lapsed*.

In using the BNC, the project used a resource that had already been automatically tokenized and part-of-speech tagged by the CLAWS tagger. For lemmatization, the project used a package kindly made available by John Carroll of the University of Sussex; see [Minnen et al. 2000]. The parser was implemented as a regular-expression pattern-matcher operating over part-of-speech tags. Thus, a simplified version of the pattern used to identify head nouns of subjects for verbs was

- The first noun encountered to the left of the verb, with any number of intervening modals, auxiliaries, adverbs, *not* and interjections.

Clearly, for many sentences, no subject was found.

Word Sketches were developed within the context of the WASPS project, which aims to develop the synergy between corpus lexicography and Word Sense Disambiguation (WSD) technology. WSD is the task of automatically finding which dictionary sense of a word applies, in a given corpus context [Ide and Veronis 1998]. In this, it takes forward work done by Clear [1994] and the HECTOR project [Atkins 1993]. In the WASPS workbench, Word Sketches serve as input to an interactive system for developing, simultaneously, an accurate analysis of the word's meaning into distinct senses and a high-precision WSD program for disambiguating it. Word Sketches and WASPS are fully described in Kilgarriff and Tugwell [2001a, 2001b].

4.2 Grammatical relations

The Word Sketch approach requires an inventory of grammatical relations, so that a collocation list can be developed for each. The inventory was identified by considering which grammatical relations often hold lexicographically interesting facts. In English, the most obvious cases include:

- for verbs: subject and object (nouns regularly occupying these positions), modifying adverbs
- for nouns: "subject-of" and "object-of" (verbs of which the noun is regularly a subject or object), modifying adjectives, other nouns appearing in compounds with the nodeword
- for adjectives: noun complements, modifying adverbs
- for all three word classes:
    - prepositional complements: a trinary relation between the nodeword, the preposition, and the content-word collocate: e.g. conversation **with** (a) **friend**
    - "and/or" relations (often a revealing set: e.g. bitter **and protracted**, bitter **and resentful**, bitter **and unpleasant**)

In addition to the binary and trinary relations, we used several unary relations such as “plural”, for nouns, and “passive” for verbs: thus, for example, the verbs **ban** and **found** both show a high proportion of passive instances, and this is reflected in the MED dictionary entries (both in the example sentences and in a grammar note "often in passive").

There will always be salient collocations which do not fit a limited list of grammatical relations. Word Sketches currently have a "fallback" procedure for identifying high-salience collocations not complying with the grammatical relations, but these suffer the drawbacks of grammatically-blind collocate lists discussed above and need further work.

4.3 Lexicographic salience statistic

The salience statistic we used in Word Sketches is the product of MI and the logarithm of the raw frequency of the collocate. We have found that MI and Log-likelihood both present too many low-frequency collocates, relative to lexicographers’ concerns, so we compensate by multiplying by the log of the frequency. The statistic lacks mathematical credentials; however it is not apparent to us that lexicographers’ needs match any well-founded mathematical model. Our empirical experience is that this statistic mediates well between the lexicographer’s wish to see high-MI collocations, and their wish to see high-frequency ones.

5. A collaborative project

In a collaboration between the University of Brighton and Bloomsbury Publishing plc, lexicographers creating the text for the new *Macmillan English Dictionary* (MED) [Rundell 2002] were supplied from the start of the project with Word Sketches for over 8000 English words – specifically, for all the most frequent nouns, verbs, and adjectives in the lexicon of English. The MED is an advanced-level monolingual learner's dictionary (MLD), based on a corpus of a little over 200 million words (of which the BNC forms the largest component). The most recent generation of MLDs has focussed on word combinations of various types [e.g. Rundell 1998. 322, 324], and one of the objectives of the MED was to supply high-quality information about common collocations in English, and to do so as systematically as possible.

From an editorial point of view, it was envisaged that the Word Sketches would provide lexicographers with a concise yet fine-grained summary of the collocational preferences of the most frequent and descriptionally-complex words in English. Furthermore, the software would give the dictionary improved claims to completeness and reduce the risk of significant behaviour patterns being missed. From the point of view of the publishing management, meanwhile, the Brighton/Bloomsbury collaboration was expected to generate significant savings in editorial time, by reducing the need for large-scale concordance scanning. And so, to a large extent, it proved.

5.1 A simple example: forge

The Word Sketch for the verb **forge** includes the following list of frequently-occurring nouns in the object position, shown in descending order of significance:

link	73
alliance	25
bond	14
partnership	11
signature	6
relationship	13
unity	6
tie	6
career	8
letter	10
friendship	4
contact	6
document	6
coupon	3
passport	3

Table 2: Extract from Word Sketch for **forge** (verb)

The software allows us to go back to the raw data in the corpus at any stage: clicking on any of the words in the list will bring up a set of corpus sentences exemplifying the specified pattern (e.g. **forge+partnership**). But before we even get that far, a quick glance at the list gives us a very clear idea of the way this verb behaves. It could be argued that relationships of this (simple) type might emerge equally clearly from a set of right-ordered concordances, but in fact this is far from being the case. When the Word Sketch counts 25 instances of the collocate **alliance**, for example, these will include corpus lines that would not appear adjacently in any concordance display, such as:

a scheme to ***forge*** *an informal* ***alliance*** *with Mr Mandela and the African National Congress*

An even more important ***alliance****, in terms of international power politics, had been* ***forged*** *in January 1964*

the General Council pursued its intention of ***forging*** *an effective industrial* ***alliance*** *between ...*

By identifying all such instances of **alliance** as objects of the nodeword, the program instantly highlights a relationship that would almost certainly have taken much longer to discern by traditional means.

5.2 Methodological implications: a case study

One of the most demanding of all the lexicographer's tasks is the process of developing a coherent schema of sense divisions for a complex lexical item. In reality, dealing with words like **forge** is not unduly challenging – though even here the time-savings offered by Word

Sketches are significant. But the most interesting outcome of this collaboration was the discovery that – for genuinely difficult words – the Word Sketches provided far more than just a rapid summary of collocational preferences.
Consider the Word Sketch for **challenge**, an extract of which is shown below[4].

Word Sketch for *challenge* (n)

BNC freq=6448, rank=1586

PP_to	793	**PP_for**	191	**object_of**	2004	**prep**	2499	**adj**	2032
	13.73:1		2.85:1		2.1:1		1.66:1		1.43:1
leadership	18	title	7	face	138	to	892	biggest -	47
authority	41	championship	5	meet	199	from	148	serious -	65
status_quo	5	place	7	pose	53	for	235	greatest -	42
auxerre	3	industry	5	present	71	against	27	intellectual -	26
dominance	5	honour	3	mount	33	of	652	legal -	50
decision	14	leadership	3	relish	17			direct -	44
expert	7			accept	58			new -	156
legitimacy	4			resist	19			daunting -	10
order	12			enjoy	35			larval -	8
legality	3			represent	36			major -	57
validity	4			issue	23			formidable -	13
Thatcher	5			constitute	17			exciting -	20
integrity	4			tackle	13			real -	44
wisdom	4			launch	16			strong -	34
orthodoxy	3			offer	31			toughest -	7
power	13			withstand	6			solar -	10
supremacy	3			provide	40			enormous -	15
rule	9			evade	5			fundamental -	15
idea	11			counter	6			blind -	12
government	16			maintain	15			environmental -	18

Table3: Extract from the Word Sketch for **challenge** (noun)

The first list here, showing nouns that frequently appear in the string "a challenge *to* –", divides fairly neatly into sets meaning "prevailing ideas" (**orthodoxy, wisdom, idea** etc) and "the prevailing power structures or power holders" (**leadership, authority, the status quo, dominance, supremacy**). And those two words **validity** and **legitimacy** more or less encapsulate the difference between these two types of **challenge**. The third column, a list of verbs of which **challenge** regularly appears as an object, divides mainly into words meaning "be presented with a challenge" (**face, meet**), "deal with a challenge" (**relish, enjoy, tackle, withstand, counter**), and "constitute a challenge" (**pose, present, represent, constitute**). The noun itself retains broadly the same meaning in all these cases, but the contextual information contributes significantly to our overall understanding of the word: for example, it emerges from items in both the verb list and the adjective list that a challenge, though an inherently difficult proposition, may be something that one can derive pleasure from dealing with. However, a fourth set of verbs here (words like **mount, issue**, and **launch**) hints at a quite separate meaning, where a human agent *initiates* a challenge – and this use is linked to the words in the first list (and indeed to most of those in the second list too). Items in the adjective list, meanwhile, relate closely to one or other of the sets identified in the other lists: for example, a click on the word **direct** takes us to a corpus line about someone "*launching* a *direct* challenge *to* male *authority*".

There is far more, of course, but this necessarily brief overview will give some idea of the diagnostic power of the Word Sketches. For what looks on the surface like a set of discrete lists, each illustrating a particular combinatorial frame, turns out to be a very compact snapshot that reveals most of the key features of a word's behaviour, and contributes critically to the process of analyzing a word's behaviour into its distinct meanings. As Sue Atkins puts it, using Word Sketches "radically reduces the time it takes to get an overview of the behaviour of the lexeme" [Atkins, this volume]. For the editorial team that created the MED, the program was initially perceived as a useful supplement to the well-established technique of scanning concordance lines, specifically for the task of identifying important collocates. Before long, however, the Word Sketches came to be the lexicographer's preferred starting point for analyzing a given word; concordance-scanning still formed an important part of the process, of course, but it was no longer the primary mode of investigation. On this project, at least, the methodology used for analyzing corpus data underwent a significant change, and this may have implications for all of us working in corpus lexicography.

Our experiences suggest that any lexicography project that can gain access to a large corpus would benefit from summarizing corpus data in Word Sketches. This is not as forbidding as it may sound. It requires some Natural Language Processing tools (lemmatizer, part-of-speech tagger, parser) but these tools are available for several languages, and where they are not available, there are several possible strategies. Lemmatizers generally implement a modest number of rules which lexicographers will already know well, so, with some computational support, a lemmatizer can be developed. For part-of-speech taggers, tools which can be trained for different languages are available (see eg http://www.xrce.xerox.com/competencies/content-analysis/fsnlp/train.html).

A parser can be readily implemented by matching patterns of part-of-speech tags. This might make many errors, but as all results are passed through the filter of the salience statistics, and lexicographers can readily ignore a measure of noise, this is not critical. Our closing note is, therefore, to say "you can do it too", and to encourage collaborations between dictionary-makers and computationalists across ever more languages.

References

[Atkins 1993] B.T. S. Atkins, 1993. Tools for computer-aided corpus lexicography: The Hector Project, in: *Acta Linguistica Hungarica* 41, pp 5-72.

[Church et al. 1991] Kenneth Church, William Gale, Patrick Hanks and Donald Hindle, 1991. Using Statistics in Lexical Analysis, in: Uri Zernik (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum, New Jersey, pp 115-164.

[Church and Hanks 1989] Kenneth Church and Patrick Hanks, 1989. Word association norms, mutual information and lexicography, in: *ACL Proceedings, 27th Annual Meeting*, Vancouver, pp 76-83.

[Clear 1994] J. Clear, 1994. I can't see the sense in a large corpus, in: *Papers in Computational Lexicography (COMPLEX).* Budapest, pp 33-48.

[Dunning 1993] Ted Dunning, 1993. Accurate methods for the statistics of surprise and coincidence, in: *Computational Linguistics* 19 (1), pp 61-74.

[Grefenstette1998] Gregory Grefenstette, 1998. The future of linguistics and lexicographers: will there be lexicograpghers in the year 3000?, in: T. Fontenelle et al. (eds.) *EURALEX 1998 Proceedings*. Liège, University of Liège, pp 25-41.

[Heid et al. 2000] Ulrich Heid, Stefan Evert, Vincent Docherty, Wolfgang Worsch and Matthias Wermke, 2000. Computational tools for semi-automatic corpus-based updating of Dictionaries, in: *EURALEX 2000*, Stuttgart, pp 183-196.

[Ide and Véronis 1998] Nancy Ide and Jean Véronis, 1998. Introduction to the special issue on word sense disambiguation: the state of the art, in: *Computational Linguistics* 24 (1), pp 1-40.

[Kilgarriff and Tugwell 2001a] Adam Kilgarriff and David Tugwell, 2001. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, in: *Proceedings of the Collocations Workshop*, ACL 2001. Toulouse, pp 32-38.

[Kilgarriff and Tugwell 2001b] Adam Kilgarriff and David Tugwell, 2001. WASP-bench: an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation, in: *Proceedings of the Machine Translation Summit VIII*, pp 187-190.

[Minnen et al. 2000] Guido Minnen, John Carroll and Darren Pearce, 2000. Robust, Applied Morphological Generation, in: *First International Conference on Natural Language Generation*. Mitzpe Ramon, Israel, pp 201-208.

[Pedersen 1996] Ted Pedersen, 1996. Fishing for Exactness, in: *Proc. Conf. South-Central SAS Users Group*. Texas.

[Rundell 1998] Michael Rundell, 1998. Recent trends in English pedagogical lexicography, in: *International Journal of Lexicography*. 11.4, pp 315-342.

[Rundell 2002] Michael Rundell (Editor-in-Chief). *The Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Publishers Limited.

Endnotes

[1] In the original proposal, T-score is used to identify words which show the difference in collocational behaviour between near-synonyms, so has three arguments: the two near-synonyms and the collocation. But lexicographic schedules rarely allow that level of delicacy of analysis, and T-score has most widely been used simply to measure noteworthiness between a word and its collocates.

[2] A partial exception is the "Picture" software used by the COBUILD team, which lists high-scoring collocates for each position relative to the keyword.

[3] See now Grefenstette 1998 for a good explanation of how each of these technologies can benefit the lexicographic process.

[4] The extract here shows the first five lists from a total of twelve for this word.