

Semi-Automatic Dictionary Drafting

Adam Kilgarriff and Pavel Rychlý

Lexical Computing Ltd, UK; Masaryk University, Brno, Czech Republic

1 The big picture

How does language work? The Fregean tradition, picking up from Aristotle and Leibniz and carried forward by Quine, Davidson and Montague, gives some glimpses of how the meanings of words and phrases combine, using grammar rules, to give meanings of sentences. Formal work on discourse and dialogue gives hope for an understanding of how the sentences build the ‘meanings’ (or, better, achieve the communicative purposes) of larger units. But what of the words and phrases? What – or, better, how- do they mean?

Here the lofty realms of philosophy and theoretical linguistics collide uncomfortably with the endless detail of lexicography. The lofty realms, all by themselves, are of no use. As Montague succinctly put it, when asked “what is the meaning of life” (I imagine the encounter on a remote Californian beach against a backdrop of sand dunes, an enthusiastic student spotting the great man at a distance, nervously approaching him and bravely interrupting his reverie): “life prime”.¹

For the last thirty years, Patrick has been forging an account of how words and phrases work that reconciles the challenge of the philosophy with the detail – the enormous, particular, sometimes concrete, sometimes abstract, sometimes banal and sometimes poetic, often bewildering and often extraordinary detail of what words do. Louis MacNeice says, “World is crazier and more of it than we think”: Patrick shows us the same can be said of word.

2 Norms and exploitations

Patrick’s Theory of Norms and Exploitations gives an account of how each word and phrase has normal uses, and how the meaning potential embedded in these normal uses can be exploited to make the word work outside those situations. The lexicographer’s job is to capture the norms. The norms can best be expressed as corpus patterns, where corpus patterns incorporate grammar, collocations and semantic categories of collocates, and entailments: what follows if we use the word in this way, or (philosophical objections aside) its meaning. Over the last decade Patrick has undertaken Corpus Pattern Analysis for several hundred of the core verbs of English.

Once a word has been exploited to do something that has not, up to that point, been normal for it, it is always possible that a community of speakers adopt the exploitation. Then the once non-normal use becomes a new norm. The question “is a word’s use metaphorical” is not an interesting one, since so many norms were once metaphorical exploitations, but have now become part of the word’s standard repertoire. The interesting question is whether a word’s use exploits the word’s potential in a novel way.

Consider Leonard Cohen’s

I followed the course
From chaos to art

Desire the horse
Depression the cart

Horse-and-cart is a clear visual image. As I read the verse for the first time, I see the large piebald carthorse plodding along with laden cart behind. The metaphor takes the pair and gives them roles in the central struggle of the artist's life. As he does it so well (and therein lies the craft of the poet), something new is added to the knowledge of *horse and cart* in the reader's mind, and the reader's uses of the expression will thereafter have the potential to exploit the reference.

2.2 Corpus Patterns

Corpus patterns for the English verb *shuffle* are as below.

1	54%	[[Human]] shuffle [NO OBJ] [Adv[Direction]] [[Human]] moves [[Direction]] slowly, without lifting the feet off the ground typically because [[Human]] is old, indecisive, or not in a hurry
2	28%	[[Human]] shuffle {{{[Document = PLURAL]}} {cards}} [[Human]] rapidly rearranges the order of {{{[Document = PLURAL]}} cards} typically [[Human]] does this to playing cards or papers by a series of deft movements of the hands
3	3%	[[Human]] shuffle [NO OBJ] {through [[Document]]} [[Human]] looks at or reads the pages of [[Document]] very quickly and superficially
4	4%	[[Human 1 = Head of Government Head of Institution]] shuffle {{cabinet ministers ...} [[Human 2 = Minister Executive = PLURAL]]} [[Human 1 = Head of Government Head of Institution]] assigns new roles to [[Human 2 = Minister Executive]], moving some from one post to another, firing others, and sometimes bringing in new recruits
5	4%	[[Human 1]] shuffle {{responsibility obligation}} {off} (to onto [[Human 2]]) [[Human 1]] refuses to have {responsibility, obligation} for some action (and passes or tries to pass {responsibility, obligation} to [[Human 2]])
6	4%	<i>idiom</i> [[Human]] shuffle {mortal coil} {off} [[Human]] dies

The patterns are as found in CPA online.ⁱⁱ The online version has in addition links to the associated concordance, and the fact that the manually-allocated sample size was 114. The entry shows that six patterns were found for the verb, each accounting for the percentage of the sample shown in the second column. The third column gives, first, the pattern, and then the implicatures of the pattern: what this use of the verb 'means'. Within the patterns, collocates are within curly brackets and semantic categories within square brackets. For further details of the formalism see the website.

The method for creating the CPA entry is for a lexicographer to study the corpus data for the word: in particular, to

- take a sample of corpus lines, initially 200 in most cases
- examine this set to exclude errors (where, for example, the headword was a mis-spelling for another word, or was in a web page of 'word salad' generated by computer, or was a noun wrongly tagged as a verb; in this case this left 114 instances). For the remainder,
- identify the patterns and allocate the instances to them, in the process drawing up the formal statements as above.

The patterns are intimately related to what, in usual lexicographic practice, would be called word meanings or word senses. There is usually a one-to-one or many-to-one mapping between patterns and senses in a good dictionary. However CPA does not work with an explicit concept of word sense or word meaning. It prefers to remain uncommitted, following Wittgenstein's advice: "don't ask for the meaning, ask for the use".

CPA patterns are not automatically derived from corpora, nor are they 'implemented': they are not designed to be used as patterns for automatic corpus searching, to find all the corpus instances matching the pattern (and no non-matches). However they are not so different from patterns which could be used in this way, for example using the corpus query language CQL.ⁱⁱⁱ The main stumbling block to them being implemented in this way is the semantic categories. Whereas word forms, lemmas, grammatical classes and word class labels can all straightforwardly be converted into components of a search pattern, and optionality and variability can be handled using regular expressions, there is no straightforward way to translate semantic category labels like Document or Human into search pattern components. (Possible methods are under consideration by both ourselves and the CPA team.) If this problem was solved, we would then be in a position to assess the recall and precision of the CPA patterns.

3 Word Senses, and the Dream of the Disambiguating Dictionary

It is now commonplace to link a dictionary to electronic texts (in word processors, web browsers or other tools) so that, by clicking or hovering over a word, the user can see the entry for the word in the dictionary. Many publishers offer their dictionaries in this form. The basic task is easy: it is one of matching the string in the text to a headword in the dictionary. Publishers have more, or less, successful solutions to the associated issues of correctly identifying dictionary headwords for the inflected forms found in texts, and of matching multi-word expressions. (There are additional difficulties for languages which do not put spaces between words.)

One thing they do not do is take the user to the correct sense of a polysemous word. This is desirable. The user would no longer need to read the whole entry and work out which sense was relevant. For long entries this can be a forbidding task, particularly for learners who are struggling with the language in the first place. If the dictionary is bilingual, then the correct sense becomes the correct translation: the user could be directly given an appropriate translation.

If the dictionary publisher had the ability to disambiguate in this way for the user, then they would also have the ability to disambiguate offline, and that would be a great boon for automatic translation and a range of other language technology applications including question answering, information retrieval and information extraction. They would have a disambiguating dictionary, and they would have solved the great problem of Word Sense Disambiguation (WSD).

WSD has been a challenge for language technology researchers since the earliest days of the field (see Agirre and Edmonds 2006 for a wide-ranging review of the field and description of the state of the art). It remains painfully intractable, with all systems in

the SENSEVAL and SEMEVAL competitions making errors in over a quarter of cases.^{iv} So the disambiguating dictionary (which does not make many, many errors) remains a dream.

But what steps might be made in that direction? In Kilgarriff (2005) we make the case that collocations provide a productive framework for thinking about the issue. Yarowsky (1993) put forward the ‘one sense per collocation’ hypothesis: to the extent that it is true, collocates^v serve to disambiguate. As a step towards disambiguating occurrences of words in running text, we can associate a word’s collocations with one or other of its senses. This will probably be an easier task than full WSD because a collocation is only a collocation if it is a reasonably common pattern of usage for a word, so we will be able to find multiple examples of it in a sufficiently large corpus, so:

- We will not be aiming to disambiguate exploitations (which would be doomed, because exploitations will not be covered by the dictionary), but only norms
- We will always have multiple contexts of a collocation to use as input to any disambiguation algorithm
- Collocations are often given (implicitly or explicitly) in dictionary entries
- It is a bounded task: whereas a word can appear in any number of contexts, its collocations will count in the tens or possibly in the hundreds.

Also, WSD systems generally work through collocations: they aim to find collocations (as well as grammatical patterns and domains) associated with each sense of the word, and then use them as clues to disambiguate new instances. So, if our goal is just to disambiguate collocations, we are ‘doing WSD’ but stopping before we attempt the most difficult part.

How then might we associate collocations with senses? There are three options: by hand, by computer, or half-and-half. To do it by hand is a very large undertaking: there are perhaps 10,000 polysemous words to be covered (Moon, 2000) and perhaps an average of twenty or thirty collocations to be assigned per sense. Fully automatic methods are possible but are likely to make many errors. Semi-automatic methods look promising, and have been tried in the WASPS project (Kilgarriff et al, 2003).

4 Semi-Automatic Dictionary Drafting (SADD)

Here we take the ideas in WASPS and update them, integrating them into the same framework and software that have been used in CPA and adding further steps of semi-automation in a prototype called SADD. If successful, it will allow the computer to do much of the footwork of CPA, and will result in corpus patterns which are semi-automatically derived and implemented: they will allow us to assign new corpus instances to senses (with recall and precision yet to be determined).

We conceptualise the problem as follows: the basic objects in our world are specific instances of words in use. A word sense is a grouping of these instances (as is a CPA pattern). When lexicographers study the corpus evidence for a word, to arrive at the set of senses that will go into the dictionary, what they are doing is grouping the corpus lines according to similarities of form and meaning. Each group is then a distinct sense.^{vi}

The challenge for SADD is then to cluster the instances as a lexicographer would. For this, we need to identify the features of each instance that correspond to the aspects of form and meaning that the lexicographer uses for their grouping. Collocation, and the grammar pattern, are the most immediately available and useful. The domain of the text, and semantic categories of collocates, are also items we would like to use though they are not immediately available.

The next challenge is how to do the clustering. The best place to start is collocations. A collocation is already a grouping of instances, and we generally get one sense per collocation. A word sketch is a very-fine-grained analysis of a word, so we can see the challenge as clustering the collocates in a word sketch. We can start the process by observing that “eat lunch” and “eat dinner” are probably the same sense of *eat* because *lunch* and *dinner* are similar.

4.1 Infrastructure: The Sketch Engine

The infrastructure we use (and which is also used in CPA) is the Sketch Engine (Kilgarriff et al 2004, <http://www.sketchengine.co.uk>). The Sketch Engine creates ‘word sketches’ – one page, corpus-driven accounts of a word’s grammatical and collocational behaviour. Word sketches have been in use in lexicography for ten years now and have received a number of positive reviews; a formal evaluation is given in Kilgarriff et al 2010. The word sketches identify the collocates that we want to assign to senses.

The Sketch Engine also prepares a distributional thesaurus (Kilgarriff and Rychly 2007). We can use the thesaurus to cluster collocates: if a headword has two collocates (in the same grammatical relation) and one is in the other’s thesaurus entry, then we put them together, as in Fig. 1.

object	58698	4.0
food 4972	11512	8.22
fish 1156 anything 790 everything 271 animal 304 heart 293 plant 298 something 448 variety 238 nothing 247 pattern 189 word 217 thing 389 place 392 quality 213 product 224 day 367 way 270 area 234		
disorder 2361	4752	9.0
diet 1385 habit 1006		
meal 1783	4334	8.32
lunch 1046 breakfast 886 dinner 619		

Fig. 1: Clustered word sketch for *eat* (verb), entry for grammatical relation OBJECT, UKWaC corpus. “Things we eat” are clustered under in the first group, under *food*; patterns of eating under *disorder*; meals under *meal*. It does not always work out as neatly as this! The numbers immediately next to the collocates are counts for that collocation and the numbers in the middle column are aggregates for all words in that cluster. The third column gives the salience for the cluster.

4.2 The Corpus

The output of a corpus tool is only as good as the corpus. The project requires a large corpus, so that, even for mid and low frequency words, we have plenty of evidence of each collocation. The language was English. The corpus we plan to use is UKWaC

(Ferraresi et al., 2008), a corpus of 1.5 billion words for English, drawn from the web, and carefully ‘cleaned’ (to remove advertisements, navigation bars, copyright statements and other ‘boilerplate’ text) and de-duplicated, with all duplicate and near-duplicate documents removed (Pomikalek, 2008).^{vii} The prototype discussed here uses the British National Corpus.^{viii}

The corpus had already been tokenized, lemmatised and part-of-speech-tagged using the leading tool TreeTagger.^{ix}

4.3 A worked example: SADD in pictures

First (in a minimal screen not shown here) we specify the word we are working on, here *charge* (verb). We then see a version of the word sketch with only the most salient collocates and collocate clusters shown, as in Fig. 2. There is a box to enter text beside each. The lexicographer examines the evidence (as they normally would), assigning a short mnemonic to each new sense they encounter: in Fig 2 the mnemonics *money*, *crime* and *electric* have been created for three senses of *charge*. The lexicographer has assigned the cluster <object, {*fee sum*}> and the collocate <pp_at-p, *rate*> to *money*, and so forth, and is in the process of assigning <subject, *magistrate*> to *crime*.

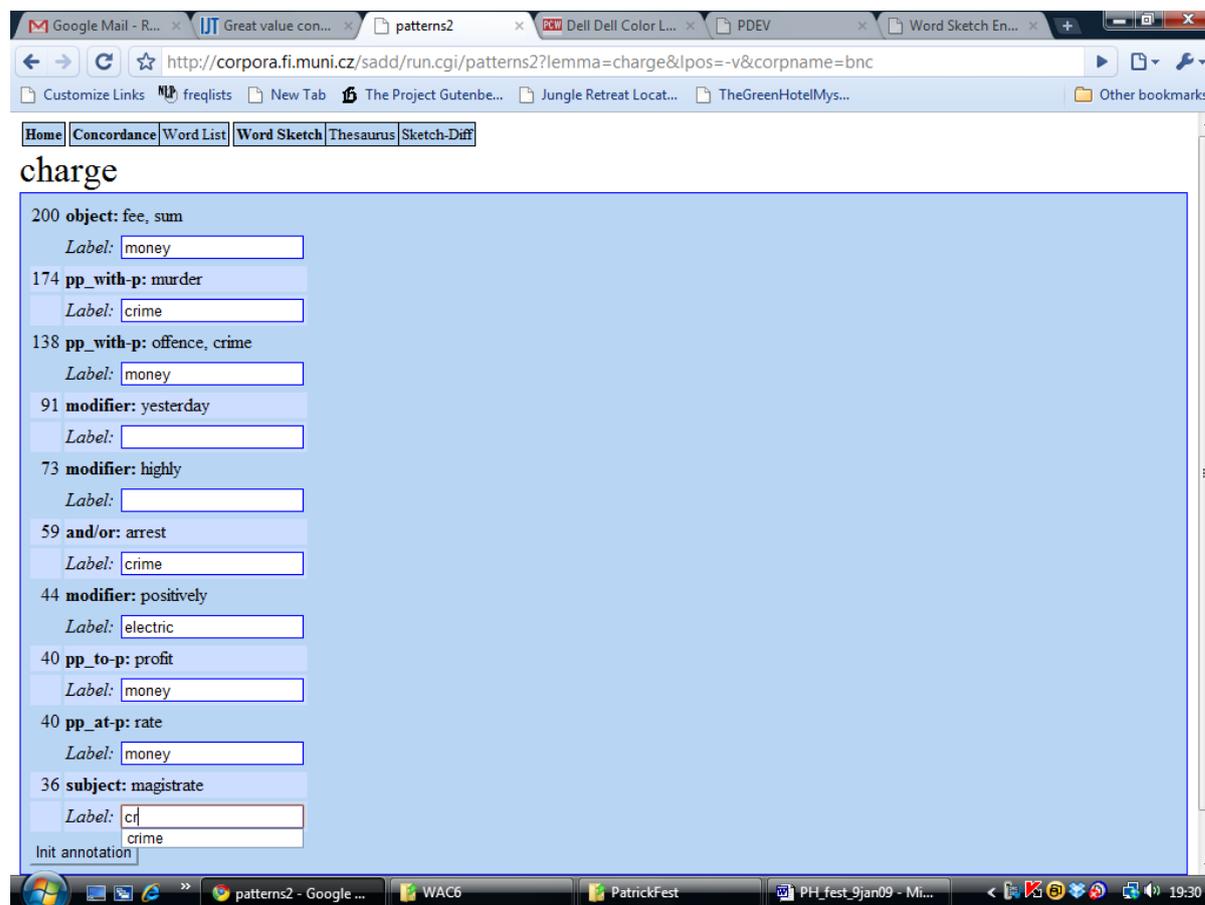


Fig. 2: Screen showing initial clustering of collocates, where the lexicographer starts the processes of assigning collocates to senses, creating a mnemonic for each sense in the process.

This process provides ‘seeds’ (Yarowsky 1995) to the WSD process. The hope is that these allocations provide enough evidence of what counts as the *money*, *crime* and *electric* sense of *charge*, for a WSD program to discover more associations by itself.

When the lexicographer clicks the ‘Init annotation’ button at the bottom of Fig. 2, this is what happens. An underlying WSD program uses the seeds to make further assignments of collocates to senses. The output, in the process for *charge*, is shown in Fig. 3. Other collocates, and the bulk of the instances, are not yet allocated.

Google Mail - R... x UJT Great value con... x pat2annot x Dell Dell Color L... x PDEV x Word Sketch En... x

http://corpora.fi.muni.cz/sadd/run.cgi/pat2annot?clseek_0=7921819,7921645,&ccllabel_0=money&clseek_1=7923328,l

Customize Links freqlists New Tab The Project Gutenbe... Jungle Retreat Locat... TheGreenHotelMys... Other bookmarks

Home Concordance Word List **Word Sketch** Thesaurus Sketch-Diff View

Corpus: **bnc**
[conc description](#)

Annotating: **charge-v**

Not assigned 6050 [P](#) / [N](#)
modifier: negatively, electrically, emotionally, *pp_with-p:* conspiracy, assault, *object:* vat, battery, *part_intrans:* in, *pp_for-p:* service, *part_trans:* up, [word sketch](#)

crime 265 [P](#) / [N](#)
pp_with-p: murder, theft, offence, *and/or:* arrest, *subject:* magistrate, *modifier:* yesterday, today, *object:* magistrate, custody, court, [word sketch](#)

electric 44 [P](#) / [N](#)
modifier: positively, *object:* nucleus, proton, atom, [word sketch](#)

money 444 [P](#) / [N](#)
object: fee, sum, person, *pp_with-p:* offence, crime, *subject:* Girobank, *pp_to-p:* profit, *pp_at-p:* rate, *modifier:* only, also, [word sketch](#)

Minscore: Mindiff: [Automatically annotate](#)



Fig. 3: Summary of current state of *charge-v* data: 6050 corpus instances have not been assigned (and the highest salience unassigned collocates are shown). For each of the three senses which have been identified we see the number of corpus instances and the assigned collocations. For each sense plus ‘Not assigned’, we can click to see the instances (P for positive), the non-instances (N for negative) and the word sketch.

We can click to see a ‘sense sketch’ for each of the senses, or of the unassigned data. The sketch for the as-yet-unassigned data is shown in Fig. 4. As can be seen, associated with each collocate is a drop-down menu where the user can select one of the mnemonics. (Other standard options, always available, are ‘u’ for error/unassignable, to include POS-tagging errors and typos; ‘x’ for exploitation – exploitations should not be assigned as they lie outside the remit of lexicography; ‘Add Sense’; and ‘None’, for undoing an erroneous assignment.)

The user also has the option of assigning individual concordance lines, in the concordance interface, using the same menu, was shown in Fig. 5. This was the interface used in CPA. It has the advantage that specific instances are assigned, so the risk of making generalisations which overlook the exceptions is avoided, but the disadvantage that it is slower: assignment by collocations implicitly assigns a whole set of instances in one operation, rather than doing one at a time. We do not yet have enough experience of using the system to be clear about the merits of the two approaches. (The first author’s initial enthusiasm for assignment-by-collocate, and

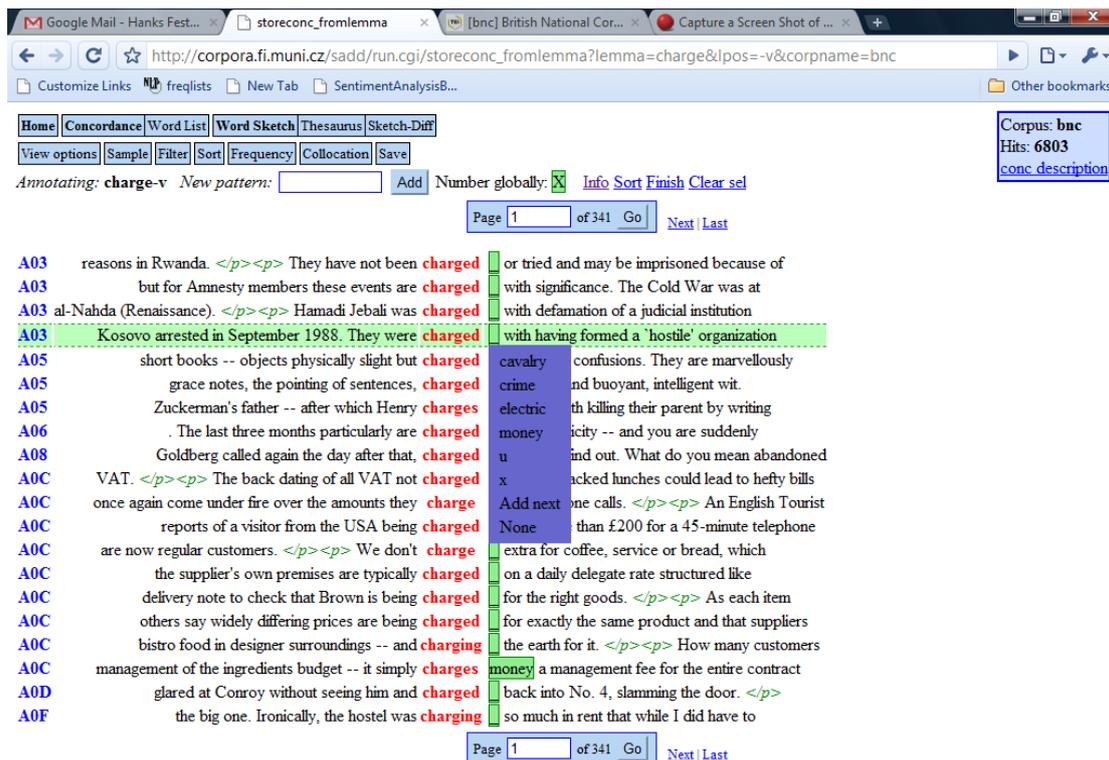


Fig. 5: Screen for assigning corpus instances to senses.

Once the user is satisfied with the detail with which allocations have been made, they have the option of running the WSD algorithm in a different mode, ‘to completion’, so that it allocates all instances of the word, where it can find any evidence, to one or other of the options. They also have the option of downloading the sense-differentiated profile for the word as XML, for loading into a dictionary editing system. At this point the method converges with ‘tickbox lexicography’ (Kilgarriff et al 2009): hence the name of the programme: we will have semi-automatically drafted the dictionary entry.

It bears reiterating that we have described a first research prototype. To turn it into a viable system for production-mode lexicography, there remains a great deal of work to be done.

5 Conclusion

In this paper we have charted how the theory of norms and exploitations makes a link between the high theory of philosophy of language, and what we find in the lexicon. We have followed Patrick in exploring what this means for the use of corpora and for lexicography, briefly describing his application of the theory in Corpus Pattern Analysis. After considering the appeal and the shortcomings of automatic word sense disambiguation, we have presented a piece of software, SADD, which aims to reconcile the excitement and enthusiasm that publishers, investors, computational linguists and software developers feel for WSD with the insights of ‘norms and exploitations’ and CPA, and the boundless messy detail in the corpus.

In J. M. Coetzee’s *Disgrace*, the protagonist is a lecturer who used to be in a department of English and his passion is, or once was, the romantic poets, but the

department has been renamed and restructured and now he teaches *Communications 101: Communication Skills* and *Communications 201: Advanced Communication Skills*. He is not happy with the statement about language that introduces them in the university handbook. ‘Human society,’ it says, ‘has created language in order that we may communicate our thoughts, feelings and intentions to each other.’ Coetzee comments:

His own opinion, which he does not air, is that the origins of speech lie in song, and the origins of song in the need to fill out with sound the overlarge and rather empty human soul.

In his writings, lectures and conversation, Patrick shows the fathomless potential that words and phrases have –in consort with us, their embodied human vehicles-- for making, breaking, layering and enriching the sum of our experience. In his account, while the role of communication is never downplayed, we see how lexis can join song in filling out the human soul.

References

- Agirre, E. and P. Edmonds (eds.) (2006) *Word Sense Disambiguation: Algorithms and Applications*. Springer.
- Agirre, E., L. Marquez, R. Wicentowski (Editors) (2009) *Language Resources and Evaluation* 43 (2), Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond.
- Ferraresi, A., E. Zanchetta, M. Baroni, S. Bernardini. (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proc. 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, Marrakech, Morocco.
- Hanks, P. Forthcoming. *Theory of Norms and Exploitations*
- Kilgarriff, A. (1997) ‘I don’t believe in word senses?’. *Computers and the Humanities* 31. Pp 97-113.
- Kilgarriff, A. (2005) Linking Dictionary and Corpus. *Proc. Asialex*, Singapore.
- Kilgarriff, A. (2006) Word Senses. In Agirre and Edmonds (eds).
- Kilgarriff, A., R. Evans, R. Koeling, M. Rundell, D. Tugwell (2003) WASPBench: a lexicographers' workstation incorporating word sense disambiguation. Demo and research note. *Proc. European ACL*. Budapest.
- Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell (2004) The Sketch Engine *Proc. Euralex*. Lorient, France: 105-116.
- Kilgarriff, A., P. Rychlý (2007) An efficient algorithm for building a distributional thesaurus *Proc. ACL Prague*.
- Kilgarriff, A., V. Kovar, P. Rychlý (2009) Tickbox Lexicography *Proc. e-Lexicography Conference* Louvain-la-Neuve, Belgium.
- Kilgarriff A., V. Kovar, S. Krek, I. Srdanovic, C. Tiberius (2010) Evaluating word sketches. *Proc. Euralex*, Leeuwarden, Netherlands.
- Krovetz, R. (1998) More than One Sense per Discourse. NEC Princeton NJ Technical Memorandum.
- Moon, R. (2000). Lexicography and Disambiguation: The Size of the Problem. In: *Computers and the Humanities* 34, p. 99-102
- Pomikálek, J. and P. Rychlý (2008) Detecting Co-Derivative Documents in Large Text Collections. In *Proc. LREC'08*. Marrakech, Morocco: 132-135.
- Schmid H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. Technical report. <http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/>
- Yarowsky, D. (1993) One sense per collocation. In *Proceedings, ARPA Human Language Technology Workshop*, pp. 266-271.
- Yarowsky, D. (1995) Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proc. 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, MA, pp. 189-196.

ⁱ A brief investigation into the source of the anecdote reveals that (1) it probably originates with not Richard Montague (not noted for his sense of humour) but, most likely, Barbara Partee, and (2) the correct written form of the response is ‘life’ though I’m not sure how to say that.

ⁱⁱ <http://deb.fi.muni.cz/pdev/> accessed on 31-12-09

ⁱⁱⁱ See eg <http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying>

^{iv} There are of course many qualifications which might be made here, as figures depend on how and what you count. Readers are referred to Agirre and Edmonds (2006) and Agirre et al (2009) for detailed discussions.

^v We use *collocation* to refer to the two-word unit, and *collocate* to refer to the word that collocates with the headword. In the Sketch Engine, collocates are identified in specific grammatical relations so collocations are word-pairs in a specific grammatical relation: a *triple* <gramrel, word1, word2>. For readability, on occasions we use *collocate* where the full version would be *collocate plus grammatical relation*.

^{vi} For the full version of this argument, see Kilgarriff (1997, 2006).

^{vii} Our version of UKWaC is 20% smaller than the one described in Ferraresi et al. (2008). Both Ferraresi's group and ours have undertaken further rounds of removing unwanted material, and shared results.

^{viii} <http://www.natcorp.ox.ac.uk>

^{ix} <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>