Adam Kilgarriff's Legacy to Computational Linguistics and Beyond

Roger Evans[1], Alexander Gelbukh[2], Gregory Grefenstette[3], Patrick Hanks[4], Miloš Jakubíček[5,6], Diana McCarthy[7](✉), Martha Palmer[8], Ted Pedersen[9], Michael Rundell[10], Pavel Rychlý[5,6], Serge Sharoff[11], and David Tugwell[12]

[1] University of Brighton, Brighton, UK
R.P.Evans@brighton.ac.uk
[2] CIC, Instituto Politécnico Nacional, Mexico City, Mexico
gelbukh@gelbukh.com
[3] IHMC, Ocala, FL, USA
ggrefenstette@ihmc.us
[4] University of Wolverhampton, Wolverhampton, UK
patrick.w.hanks@gmail.com
[5] Lexical Computing, Brighton, UK
milos.jakubicek@sketchengine.co.uk, pary@fi.muni.cz
[6] Masaryk University, Brno, Czech Republic
[7] DTAL University of Cambridge, Cambridge, UK
diana@dianamccarthy.co.uk
[8] University of Colorado, Boulder, USA
martha.palmer@colorado.edu
[9] University of Minnesota, Minneapolis, USA
tpederse@d.umn.edu
[10] Lexicography MasterClass, Brighton, UK
michael.rundell@lexmasterclass.com
[11] University of Leeds, Leeds, UK
s.sharoff@leeds.ac.uk
[12] Independent Researcher, Edinburgh, UK

Abstract. The 2016 CICLing conference was dedicated to the memory of Adam Kilgarriff who died the year before. Adam leaves behind a tremendous scientific legacy and those working in computational linguistics, other fields of linguistics and lexicography are indebted to him. This paper is a summary review of some of Adam's main scientific contributions. It is not and cannot be exhaustive. It is written by only a small selection of his large network of collaborators. Nevertheless we hope this will provide a useful summary for readers wanting to know more about the origins of work, events and software that are so widely relied upon by scientists today, and undoubtedly will continue to be so in the foreseeable future.

A. Gelbukh (Ed.): CICLing 2016, LNCS 9623, pp. 3–25, 2018.
https://doi.org/10.1007/978-3-319-75477-2_1

1 Introduction

The year 2015 was marred by the loss of Adam Kilgarriff who during the last 27 years of his life contributed greatly to the field of computational linguistics[1], as well as to other fields of linguistics and to lexicography. This paper provides a review of some of the key scientific contributions he made. His legacy is impressive, not simply in terms of the numerous academic papers, which are widely cited in many fields, but also the many scientific events and communities he founded and fostered and the commercial **Sketch Engine** software. The Sketch Engine has provided computational linguistics tools and corpora to scientists in other fields, notably lexicography for example [17,50,61], as well as facilitating research in other areas of linguistics [11,12,54,56] and our own subfield of computational linguistics [60,74].

Adam was hugely interested in lexicography from the very inception of his postgraduate career. His DPhil[2] on polysemy and subsequent interest in word sense disambiguation (WSD) and its evaluation was firmly rooted in examining corpus data and dictionary senses with a keen eye on the lexicographic process [20]. After his DPhil, Adam spent several years as a computational linguist advising Longman Dictionaries on the use of language engineering for the development of lexical databases, and he continued this line of knowledge transfer in consultancies with other publishers until realizing the potential of computational linguistics with the development of his commercial software, the Sketch Engine. The origins of this software lay in his earlier ideas of using computational linguistics tools for providing word profiles from corpus data.

For Adam, data was key. He fully appreciated the need for empirical approaches to both computational linguistics and lexicography. In computational linguistics from the 90s onwards there was a huge swing from symbolic to statistical approaches, however the choice of input data, in composition and size, was often overlooked in favor of a focus on algorithms. Furthermore, early on in this statistical tsunami, issues of replicability were not always appreciated. A large portion of his work was devoted to these issues, in his work on WSD evaluation and in his work on building and comparing corpora. His signature company slogan was ‘corpora for all’.

This paper has come together from a small sample of the very large pool of Adam’s collaborators. The sections have been written by different subsets of the authors and with different perspectives on Adam’s work and on his ideas. We hope that this approach will give the reader an overview of some of Adam’s main scientific contributions to both academia and the commercial world, while not detracting too greatly from the coherence of the article.

The article is structured as follows. Section 2 outlines Adam’s thesis and origins of his thoughts on word senses and lexicography. Section 3 continues with his

[1] In this paper, natural language processing (NLP) is used synonymously with computational linguistics.

[2] Like Oxford, the University of Sussex, where Adam undertook his doctoral training, uses **DPhil** rather than **PhD** as the abbreviation for its doctoral degrees.

subsequent work on WSD evaluation in the **Senseval** series as well as discussing his qualms about the adoption of dictionary senses in computational linguistics as an act of faith without a specific purpose in mind. Section 4 summarizes his early work on using corpus data to provide word profiles in a project known as the **WASP-bench**, the precursor to his company's[3] commercial software, the Sketch Engine. Corpus data lay at the very heart of these word profiles, and indeed just about all of Computational Linguistics from the mid 90s on. Section 5 discuss Adam's ideas for building and comparing corpus data, while Sect. 6 describes the Sketch Engine itself. Finally Sect. 7 details some of the impact Adam has had transferring ideas from computational and corpus linguistics to the field of lexicography.

2 Adam's Doctoral Research

To lay the foundation for an understanding of Adam's contribution to our field, an obvious place to start is his DPhil thesis [19]. But let us first sketch out the background in which he undertook his doctoral research. Having obtained first class honours in Philsophy and Engineering at Cambridge in 1982, Adam had spent a few years away from academia before arriving at Sussex in 1987 to undertake the Masters in Intelligent Knowledge-Based Systems, a programme which aimed to give non-computer scientists a grounding in Cognitive Science and Artificial Intelligence. This course introduced him to **Natural Language Processing** (NLP) and lexical semantics, and in 1988 he enrolled on the DPhil program, supervised by Gerald Gazdar and Roger Evans. At that time, NLP had moved away from its roots in Artificial Intelligence towards more formal approaches, with increasing interest in formal lexical issues and more elaborate models of lexical structure, such as Copestake's LKB [6] and Evans and Gazdar's DATR [9]. In addition, the idea of improving lexical coverage by exploiting digitized versions of dictionaries was gaining currency, although the advent of large-scale corpus-based approaches was still some way off. In this context, Adam set out to explore **Polysemy**, or as he put it himself:

> *What does it mean to say a word has several meanings? On what grounds do lexicographers make their judgments about the number of meanings a word has? How do the senses a dictionary lists relate to the full range of ways a word might get used? How might NLP systems deal with multiple meanings?* [19, p. 1]

Two further quotes from Adam's thesis neatly summarize the broad interdisciplinarity which characterized his approach to his thesis, and throughout his research career. The first is from the Preface:

[3] The company he founded is **Lexical Computing Ltd.** He was also a partner – with Sue Atkins and Michael Rundell – in another company, **Lexicography MasterClass**, which provides consultancy and training and runs the **Lexicom** workshops in lexicography and lexical computing; http://www.lexmasterclass.com/.

> *There are four kinds of thesis in cognitive science: formal, empirical, program-based and discursive. What sort was mine to be? ... I look round in delight to find [my thesis] does a little bit of all the things a cognitive science thesis might do!* [19, p. 6]

while the second is in the introduction to his discussion of methodology:

> *We take the study of the lexicon to be intimately related to the study of the mind ... . For an understanding of the lexicon, the contributing disciplines are lexicography, psycholinguistics and theoretical, computational and corpus linguistics.* [19, p. 4]

The distinctions made in the first of these quotes provide a neat framework to discuss the content of the thesis in more detail. Adam's own starting point was *empirical*: in two studies he demonstrated first that the range and type of sense distinctions found in a typical dictionary defied any simple systematic classification, and second that the so-called **bank model** of word senses (where senses from dictionaries were considered to be distinct and easy to enumerate and match to textual instances) did not in general reflect actual dictionary sense distinctions (which tend to overlap). A key practical consequence of this is that the then-current NLP WSD systems which assumed the bank model could never achieve the highest levels of performance in sense matching tasks.

From this practical exploration, Adam moved to more *discursive* territory. He explored the basis on which lexicographers decide which sense distinctions appear in dictionaries, and introduced an informal criterion to characterize it – the **Sufficiently Frequent and Insufficiently Predictable** (SFIP) condition, which essentially favors senses which are both common and non-obvious. However he noted that while this criterion had empirical validity as a way of circumscribing polysemy in dictionaries, it did not offer any clear understanding of the nature of polysemy itself. He argued that this is because polysemy is not a 'natural kind' but rather a cover term for several other more specific but distinct phenomena: homonymy (the bank model), alternation (systematic usage differences), collocation (lexically contextualized usage) and analogy.

This characterization led into the *formal/program-based* contribution (which in the spirit of logic-based programming paradigms collapse into one) of his thesis, for which he developed two formal descriptions of lexical alternations using the inheritance-based lexical description language DATR. His aim was to demonstrate that while on first sight much of the evidence surrounding polysemy seemed unruly and arbitrary, it was nevertheless possible, with a sufficiently expressive formal language, to characterize substantial aspects of the problem in a formal, computationally tractable way.

Adam's own summary of the key contributions of his work were typically succinct:

> *The thesis makes three principal claims, one empirical, one theoretical, and one formal and computational. The first is that the Bank Model is fatally flawed. The second is that polysemy is a concept at a crossroads, which*

> *must be understood in terms of its relation to homonymy, alternations, collocations and analogy. The third is that many of the phenomena falling under the name of polysemy can be given a concise formal description in a manner … which is well-suited to computational applications.* [19, p. 8]

With the benefit of hindsight, we can see in this thesis many of the key ideas which Adam developed over his career. In particular, the beginnings of his empirical, usage-based approach to understanding lexical behaviour, his interest in lexicography and support for the lexicographic process, and his ideas for improving the methodology and development of computational WSD systems probably first came together as the identifiable start of his subsequent journey in [20], and will all feature prominently in the remainder of this review.

What is perhaps more surprising from our present perspective is his advocacy of formal approaches to achieve some of these goals, in particular relating to NLP. Of course, in part this is just a consequence of the times and environment (and supervisory team) of his doctoral study. But while Adam was later in the forefront of lexicographic techniques based on statistical machine learning rather than formal modeling, he still retained an interest in formalizing the structure of lexical knowledge, for example in his contributions to the development of a formal mark-up scheme for dictionary entries as part of the Text Encoding Initiative [8,14,15].

3 Word Sense Disambiguation Evaluation: SENSEVAL

3.1 The Birth of Senseval98

After the years spent studying polysemy, no one understood the complexity and richness of word meanings better than Adam [62]. He looked askance at the NLP community's desire to reduce word meaning to a straight-forward classification problem, as though labeling a word in a sentence with "sense2" offered a complete solution. At the 1997 SIGLEX Workshop organised by Martha Palmer and Mark Light, which used working sessions to focus on determining appropriate evaluation techniques, Adam was a key figure, and strongly influenced the eventual plan for evaluation that gave birth to the **Senseval98** workshop, co-organized by Adam and Martha Palmer. The consensus the workshop participants came to at this meeting were clearly summarized in [24]. During the working session Adam went to great pains to explain to the participants the limitations of dictionary entries and the importance of choosing the right sense inventory, a view for which he was already well known [21,22,25]. This is well in line with the rule of thumb for all supervised machine learning: the better the original labeling, the better the resulting systems. Where word senses were concerned, it had previously not been clearly understood that the sense inventory is the key to the labeling process. This belief also prompted Adam's focus on introducing **Hector** [1] as the sense inventory for Senseval98 [43]. Although Hector covered only a subset of English vocabulary, the entries had been developed by using a corpus-based approach to produce traditional hierarchical dictionary definitions including detailed, informative descriptions of each sense [44].

This focus on high quality annotation extended to Adam's commitment to not just high **inter–tagger agreement** (ITA) but also **replicability**. Replicability measures the agreement rate between two separate teams, each of 3 annotators, who perform double-blind annotation with the third annotator adjudicating. After the tagging for Senseval98 was completed, Adam went back and measured replicability for 4 of the lexical items, achieving a staggering 95.5% agreement rate. Inter-annotator agreement of over 80% for all the tagged training data was also achieved. Adam's approach allowed for discussion and revision of ambiguities in lexical entries before tagging the final test data and calculating the ITA.

Senseval98 demonstrated to the community that there was still substantial room for improvement in the production of annotations for WSD, and spawned a second and then a third Senseval, now known as **Senseval2** [64] and **Senseval3** [59], and Senseval98 is now **Senseval1**. There was a striking difference between the ITA for Senseval98 [26,43], and the ITA for WordNet lexical entries for Senseval2, tagged by Palmer's team at Penn, which was only 71% for verbs. The carefully crafted Hector entries made a substantial difference. With lower ITA, the best system performance on the Senseval2 data was only 64%. When closely related senses were grouped together into more coarse grained senses, the ITA improved to 82%, and the system performance rose a similar amount. By the end of Senseval2 we were all converts to Adam's views on the crucial importance of sense inventories, and especially on full descriptions of each lexical entry. As the community began applying the same methodology to other semantic annotation tasks, the name was changed to **SemEval**, and the series of SemEval workshops for fostering work in semantic representations continues to this day.

3.2 Are Word Senses Real?

Adam was always thought provoking, and relished starting a good debate whenever (and however) he could. He sometimes achieved this by making rather stark and provocative statements which were intended to initiate those discussions, but in the end did not represent his actual position, which was nearly always far more nuanced. Perhaps the best example of this is his article "'I don't believe in word senses",'[4] [25]. Could a title be any more stark? If you stopped reading at the title, you would understand this to mean that Adam did not believe in word senses.[5] But of course it was never nearly that simple.[6]

Adam worked very hard to connect WSD to the art and practice of lexicography. This was important in that it made it clear that WSD really couldn't be

[4] This paper is perhaps Adam's most influential piece, having been reprinted in three different collections since its original publication.

[5] The implication that Adam *did* believe in "word senses" is controversial. There are co-authors of this article in disagreement about Adam's beliefs on word senses. Whatever Adam's beliefs were, we are indebted to him for amplifying the debate [13, 30] and for opening our eyes to other possibilities.

[6] In fact, the title is a quote which Adam attributes to Sue Atkins.

treated as yet another classification task. Adam pointed out that our notion of word senses had very much been shaped by the conventions of printed dictionaries, but that dictionary makers are driven by many practical concerns that have little to do with the philosophical and linguistic foundations of meaning and sense. While consumers have come to expect dictionaries to provide a finite list of discrete senses, Adam argued that this model is not only demonstrably false, it is overly limiting to NLP.

In reality then, what Adam did not believe in were word senses *as typically enumerated in dictionaries*. He also did not believe that word senses should be viewed as atomic units of meaning. Rather, it was the multiple occurrences of a word in context that finally revealed the sense of a word. His actual view about word senses is neatly summarized in the article's abstract, where he writes '... word senses exist only relative to a task.' Word senses are dynamic, and have to be interpreted with respect to the task at hand.

3.3 Data, Data and More Data

This emphasis on the importance of context guided his vision for leveraging corpora to better inform lexicographers' models of word senses. One of the main advantages of Hector was its close tie to examples from data, and this desire to facilitate data-driven approaches continued to motivate Adam's research. It is currently very effectively embodied in DANTE [35] and in the Sketch Engine [48].[7] This unquenchable thirst for data also led to Adam's participation in the formation of the Special Interest Group on the Web as Corpus (see Sect. 5). Where better to find endless amounts of freely available text than the World Wide Web?

4 Word Sketches

One of the key intellectual legacies of Adam's body of research is the notion that compiling sophisticated statistical profiles of word usage could form the basis of a tractable and useful bridge between corpus data (concrete and available) and linguistic conceptions of word senses (ephemeral and contentious). We refer to such profiles now as **word sketches**, a term which first appeared in papers around 2001 (for example [49,73]), but their roots go back several years earlier.

Following the completion of his doctoral thesis, Adam worked for three years at Longmans dictionary publishers, contributing to the design of their new dictionary database technology. In 1995, he returned to academia, at the University of Brighton, on a project which aimed to develop techniques to enhance (these days we might say 'enrich') automatically-acquired lexical resources. With his thesis research and lexicographic background, Adam quickly identified WSD as a critical key focus for this research, writing:

[7] The Sketch Engine, described in Sect. 6, in particular is an incredibly valuable resource that is used regularly at Colorado for revising English VerbNet class memberships and developing PropBank frame files for several languages.

> *Our hypothesis is that most NLP applications do not need to disambiguate most words that are ambiguous in a general dictionary; for those they do need to disambiguate, it would be foolish to assume that the senses to be disambiguated between correspond to those in any existing resource; and that identifying, and providing the means to resolve, the salient ambiguities will be a large part of the customization effort for any NLP application-building team. Assuming this is confirmed by the preliminary research, the tool we would provide would be for computer-aided computational lexicography. Where the person doing the customization identified a word with a salient sense-distinction, the tool would help him/her elicit (from an application-specific corpus) the contextual clues which would enable the NLP application to identify which sense applied.* [Kilgarriff 1995, personal communication to Roger Evans]

This is probably the earliest description of a tool which would eventually become the Sketch Engine (see Sect. 6), some eight years later.

The project followed a line of research in pursuit of this goal, building on Adam's thoughts on usage-based approaches to understanding lexical behavior, methodology for WSD, and his interest in support for the lexicographic process. The idea was to use parsed corpus data to provide profiles of words in terms of their collocational and syntactic behavior, for example predicate argument structure and slot fillers. This would provide a one page summary[8] of a word's behavior for lexicographers making decisions on word entries based on frequency and predictability. Crucially the software would allow users to switch seamlessly between the word sketch summary and the underlying corpus examples [73].

Adam's ideas on WSD methodology were inspired by Yarowky's 'one sense per collocation' [75] and bootstrapping approach [76]. The bootstrapping approach uses a few seed collocations, or manual labels, for sense distinctions that are relevant to the task at hand. Examples from the corpus that can be labeled with these few collocations are used as an initial set of training data. The system iteratively finds and labels more data from which further sense specific collocations are learned, thereby bootstrapping to extend coverage. Full coverage is achieved by additional heuristics such as 'one sense per document' [10]. Adam appreciated that this approach could be used with a standard WSD data set with a fixed sense inventory [49] but importantly also allow one to define the senses pertinent to the task at hand [46].

As well as the core analytic technology at the heart of the creation of word sketches, Adam always had two much more practical concerns in mind in the development of this approach: the need to deliver effective visualisation and manipulation tools for use by lexicographers (and others), and the need to develop technology that was truly scalable to handle very large corpus resources. His earliest experiments focused primarily on the analytic approach; the key deliverable of the follow-on project, WASPS, was the WASP-bench, a tool which combined off-line compilation of word-sketches with a web-based

[8] See Fig. 1, below.

interface exploring the sketches and underlying concordance data that supports them [38]; and the practical (and technological) culmination of this project is, of course, the Sketch Engine (see Sect. 6), with its interactive web interface and very large scale corpus resources.

5 Corpus Development

Understanding the data you work with was the key for Adam. As lexicography and NLP became more corpus-based, their appetite for access to more data seemed inexhaustible. Banko and Brill expressed the view that getting more data always leads to improved performance of tasks such as WSD [3]. Adam had a more nuanced view. On the one hand, he was very much in favor of using as much data as possible, hence his interest to using the power of the Web. On the other hand, he also emphasized the importance of understanding what is under the hood of a large corpus: rather than stating bluntly that my corpus is bigger than yours, a more interesting question is *how* my corpus differs from yours.

5.1 Web as Corpus

Adam's work as a lexicographer came from the corpus-based tradition to dictionary building initiated by Sue Atkins and John Sinclair with their COBUILD project, developing large corpora as a basis for lexicographic work for the Collins Dictionary. To support this work, they created progressively larger corpora of English text, culminating in the 100 million word **British National Corpus** (BNC) [53]. This corpus was designed to cover a wide variety of spoken (10%) and written (90%) 20th century British language use. It was composed of 4124 files, each tagged with a domain (informative, 75%; or imaginative, 25%), a medium tag, and a date (mostly post 1975). The philosophy behind this corpus was that it was representative of English language use, and that it could thus be used to illustrate or explain word meanings. McEnery and Wilson [58] said the word 'corpus' for lexicographic use had, at that time, four connotations: 'sampling and representativeness, finite size, machine-readable form, a standard reference.' In other words, a corpus was a disciplined, well understood, and curated source of language use.

Even though 100 million words seemed like an enormous sample, lexicographers found that the BNC missed some of the natural intuitions that they had about language use. We remember Sue Atkins mentioning in a talk that you could not discover that apples were crisp from the BNC, though she felt that *crisp* would be tightly associated with *apple*. Adam realized in the late 1990s that the newly developing World Wide Web gave access to much larger and useful samples of text than any group of people could curate, and argued [39] that we should reclaim the word 'corpus' from McEnery and Wilson's 'connotations', and consider the Web as a corpus even though it is neither finite, in a practical sense, nor a standard reference.

Early web crawlers, such as the now defunct **Altavista** engine, provided exact counts of words and phrases found in their web index. These could be used to predict the individual corpus sizes of given languages, and showed that the Web contained many orders of magnitude more text than the BNC. Language identifiers could distinguish the language of a text with a high rate of accuracy. And the Web was an open source, from which one could crawl and collect seemingly unlimited amount of text. Recognizing the limitations of statistics output by these search engines because the data on which they were based would fluctuate and thus make any experimental conditions impossible to repeat, Adam coined a phrase: 'Googleology is bad science' [31].

So he led the way in exploiting these characteristics of the Web for corpus-related work, gathering together research on recent applications of using the Web as a corpus in a special issue of *Computational Linguistics* [39], and organizing, with Marco Baroni and Silvia Bernardini, a series of **Web as Corpus** (WAC) workshops illustrating this usefulness, starting with workshops in Forli and Birmingham[9] in 2005, and then in Trento in 2006. These workshops presented initial work on building new corpora via web crawling, on creating search engines for corpus building, on detecting genres and types, and annotating web corpora, in other words, recovering some of the connotations of corpora mentioned by McEnery and Wilson. This work is illustrated by such tools at WebBootCat [5], an online tool for bootstrapping a corpus of text, given a set of seed words, part of the Sketch Engine package, of which more is said below.

Initially the multilingual collection of the Sketch Engine was populated with a range of web corpora (I-XX, where XX is AR, FR, RU, etc.), which were produced by making thousands of queries consisting of general words [70]. Later on, this was enriched with crawled corpora in the TenTen family [63], with the aim of exceeding the size of 10^{10} words per corpus crawled for each language.

One of the issues with getting corpora from the Web is the difficulty in separating what is a normal text, which you can expect in a corpus like the BNC, from what is boilerplate, i.e., navigation, layout or informational elements, which do not contribute to running text. This led to another shared task initiated by Adam, namely on evaluation of Web page cleaning. **CleanEval**[10] was a 'competitive evaluation on the topic of cleaning arbitrary web pages, with the goal of preparing web data for use as a corpus, for linguistic and language technology research and development' [55].

Under the impetus of Adam, Marco Baroni and others, Web as Corpus became a special interest group of the ACL, SIGWAC[11] and has continued to organize WAC workshops yearly. The 2016 version, WAC-X, was held in Berlin, in August 2016. Thanks to Adam's vision, lexicography broke away from a limited, curated view of corpus validation of human intuition, and has embraced a computational approach to building corpora from the Web, and using this new corpus evidence as the source for building human and machine-oriented lexicons.

[9] Working papers can be found online at http://wackybook.sslmit.unibo.it.

[10] http://cleaneval.sigwac.org.uk/.

[11] https://sigwac.org.uk/.

5.2 Corpus Analysis

The side of Adam's research reported above shows that the Web can be indeed turned into a huge corpus. However, once we started mining corpora from the Web, the next natural question is to assess their similarity to existing resources. Adam was one of the first who addressed this issue by stating 'There is a void in the heart of corpus linguistics' by referring to the lack of measures of corpus **similarity** and **homogeneity** [27]. His answer was to develop methods to show which corpora are closer to each other or which parts of corpora are similar. A frequency list can be produced for each corpus or a part of the corpus, usually in the form of lemmas or lower-cased word forms. Adam suggested two methods, one based on comparing the ranks in those frequency lists using rank statistics, such as Mann-Whitney [27], the other by using **SimpleMaths**, the ratio of frequencies regularized with an indicator of 'commonness' [33].

An extension of this research was his interest in measuring how good a corpus is. We can easily mine multiple corpora from the Web using different methods and for different purposes. Extrinsic evaluation of a corpus might be performed in a number of ways. For example, Adam suggested a lexicographic task: how good a corpus is for extraction of collocates on the grounds that a more homogeneous corpus is likely to produce more useful collocates given the same set of methods [47].

The frequency lists are good not only for the task of comparing corpora. They were recognized early on as one of the useful outputs of corpus linguistics: for pedagogical applications it is important to know which words are more common, so that they can be introduced earlier. Adam's statistical work with the BNC led to the popularity of his BNC frequency list, which has been used in defining the English Language Teaching curriculum in Japan.[12] However, he also realized the limitations of uncritical applications of the frequency lists by formulating the 'whelks' and 'banana' problems [22]. *Whelk* is a relatively infrequent word in English. However, if a text is about whelks, this word is likely to be used in nearly every sentence. In the end, this can considerably elevate its position in a frequency list, even if this word is used only in a small number of texts. Words like *banana* present an opposite problem: no matter how many times in our daily lives we operate with everyday objects and how important they are for the learners, we do not necessarily refer to them in texts we write. Therefore, their position in the frequency lists can become quite low [34], this also explains the *crisp apples* case mentioned above. The 'banana' problem was addressed in the Kelly project by balancing the frequency lists for a language through translation: a word is more important for the learners if it appears as a translation in several frequency lists obtained from comparable corpora in different languages [37].

6 The Sketch Engine

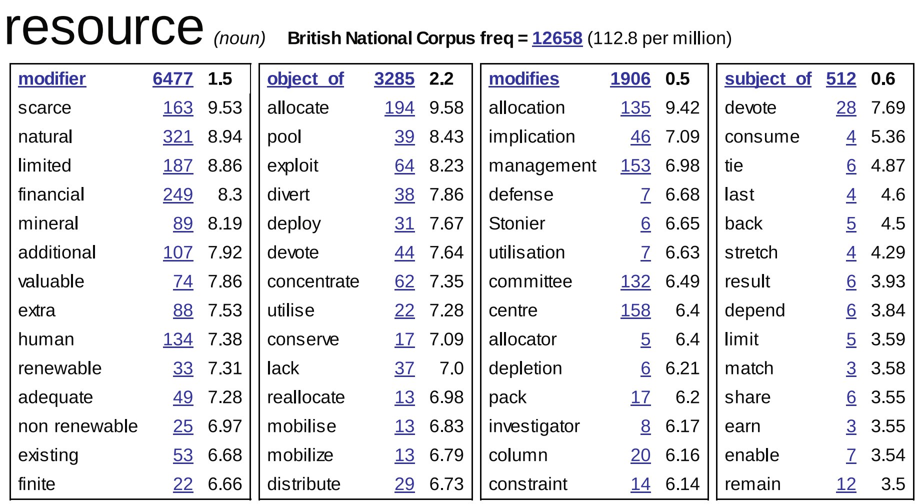
In terms of practical applications – software – Adam' main legacy is undoubtedly the Sketch Engine [36,48]: a corpus management platform hosting by 2016 hun-

[12] Personal communication from Adam to Serge Sharoff.

dreds of preloaded corpora in over 80 languages and allowing users to easily build new corpora either from their own texts or using method like the aforementioned WebBootCAT.

Sketch Engine has two fathers: in 2002 Adam' Kilgarriff met at a workshop in Brighton Pavel Rychlý, a computer scientist developing at the time a simple concordancer (Bonito [68]) based on its own database backbone devised solely for the purposes of corpus indexing (Manatee [67]).

This meeting was pivotal: Adam, fascinated by language and the potential of corpus-enabled NLP methods for lexicography and elsewhere, and looking for somebody to implement his word profiling methodology (see Sect. 4) on a large scale; and Pavel, the not-so-fascinated-by-language but eager to find out how to solve all the computationally interesting tasks corpus processing has brought in an effective manner.



resource *(noun)* **British National Corpus freq = 12658** (112.8 per million)

modifier	6477	1.5	object_of	3285	2.2	modifies	1906	0.5	subject_of	512	0.6
scarce	163	9.53	allocate	194	9.58	allocation	135	9.42	devote	28	7.69
natural	321	8.94	pool	39	8.43	implication	46	7.09	consume	4	5.36
limited	187	8.86	exploit	64	8.23	management	153	6.98	tie	6	4.87
financial	249	8.3	divert	38	7.86	defense	7	6.68	last	4	4.6
mineral	89	8.19	deploy	31	7.67	Stonier	6	6.65	back	5	4.5
additional	107	7.92	devote	44	7.64	utilisation	7	6.63	stretch	4	4.29
valuable	74	7.86	concentrate	62	7.35	committee	132	6.49	result	6	3.93
extra	88	7.53	utilise	22	7.28	centre	158	6.4	depend	6	3.84
human	134	7.38	conserve	17	7.09	allocator	5	6.4	limit	5	3.59
renewable	33	7.31	lack	37	7.0	depletion	6	6.21	match	3	3.58
adequate	49	7.28	reallocate	13	6.98	pack	17	6.2	share	6	3.55
non renewable	25	6.97	mobilise	13	6.83	investigator	8	6.17	earn	3	3.55
existing	53	6.68	mobilize	13	6.79	column	20	6.16	enable	7	3.54
finite	22	6.66	distribute	29	6.73	constraint	14	6.14	remain	12	3.5

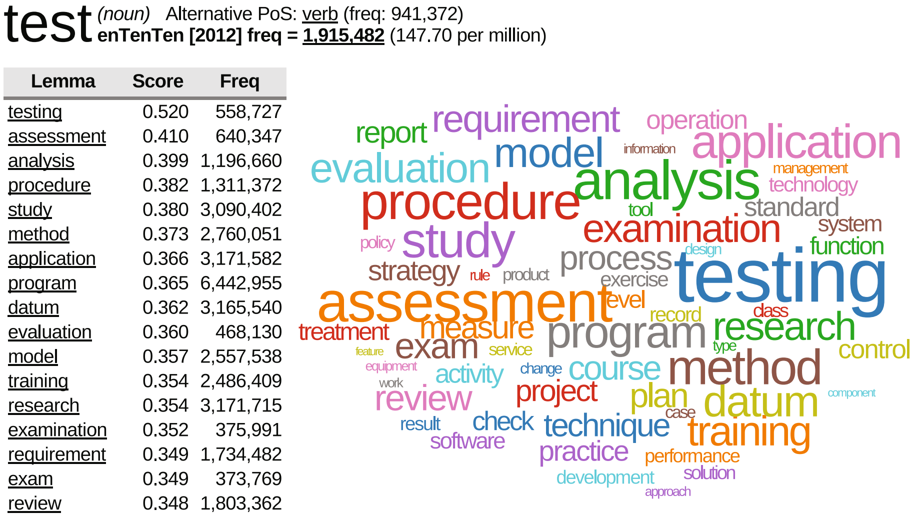
Fig. 1. Example word sketch table for the English noun *resource* from the British National Corpus.

A year later the Sketch Engine was born, at the time being pretty much the Bonito concordancer enhanced with word sketches for English. The tool quickly gained a good reputation and was adopted by major British publishing houses, allowing sustainable maintenance and – fortunately for the computational linguistics community – Adam, a researcher dressed as businessman, reinvested all company income always into further development. In a recent survey among European lexicographers the Sketch Engine was their most used corpus query system [52].

Now 12 years later the Sketch Engine offers a wide range of corpus analysis functions on top of billion-word corpora for many language and tries to fulfill Adam's goal of 'Corpora for all' and 'Bringing corpora to the masses'.

Besides lexicographers and linguists (which now implies corpus linguists — almost always) this attracts teachers (not only at universities), students, language learners, and more increasingly translators, terminologists or copywriters.

The name Sketch Engine originates from the system's key function: word sketches, one page summaries of a word's collocational behavior in particular grammatical relations (see Fig. 1). Word sketches are computed by evaluating a set of corpus queries (called the **word sketch grammar**; see [16]) that generate a very large set of headword-collocation candidate pairs together with links to their particular occurrences in the corpus. Next, each collocation candidate is scored using a lexicographic association measure (in this case, the logDice measure [69]) and displayed in the word sketch table sorted by this score (or, alternatively, by raw frequency).



test *(noun)* Alternative PoS: verb (freq: 941,372)
enTenTen [2012] freq = 1,915,482 (147.70 per million)

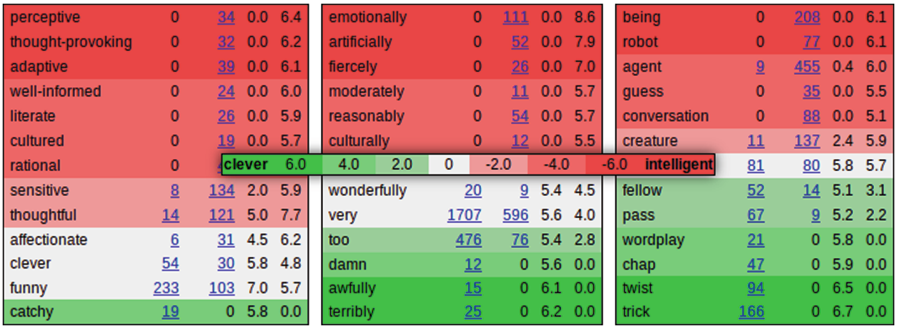
Lemma	Score	Freq
testing	0.520	558,727
assessment	0.410	640,347
analysis	0.399	1,196,660
procedure	0.382	1,311,372
study	0.380	3,090,402
method	0.373	2,760,051
application	0.366	3,171,582
program	0.365	6,442,955
datum	0.362	3,165,540
evaluation	0.360	468,130
model	0.357	2,557,538
training	0.354	2,486,409
research	0.354	3,171,715
examination	0.352	375,991
requirement	0.349	1,734,482
exam	0.349	373,769
review	0.348	1,803,362

Fig. 2. Thesaurus entry for the English noun *test* computed from the word sketch database generated from the enTenTen12 corpus.

This hybrid approach – a combination of handcrafted language-specific grammar rules with a simple language independent statistical measure – has proved to be very robust with regard to the noisy web corpora, and very scalable so as to be able to benefit from their large size. Further on, devising a new sketch grammar for another language turned out to be mostly a straightforward task, and usually a matter of a few days of joint work between an informed native speaker and somebody who is familiar with the corpus query language. Sketch grammars can be adapted for many purposes, for example a recent adaptation incorporated automatic semantic annotations of the predicate argument fillers [57]. As of 2016 the Sketch Engine contains word sketch grammars for 26 languages, and new ones are being added regularly. In addition, the same formalism has been successfully used to identify key terms using Adam's SimpleMaths methodology [41].

Two additional features are also provided building on the word sketches: a distributional thesaurus and a word sketch comparison for two headwords called sketch-diff. The distributional thesaurus is computed from the word sketch index by identifying the most common words that co-occur with the same words in the same grammatical relations as a given input headword. Therefore the result is a set of synonyms, antonyms, hyper- and hyponyms — all kinds of semantically related words (see Fig. 2).

The sketch difference identifies the most different collocations for two input headwords (or the same headword in two different subcorpora) by subtracting the word sketch scores for all collocations of both headwords (see Fig. 3).



perceptive	0	34	0.0	6.4
thought-provoking	0	32	0.0	6.2
adaptive	0	39	0.0	6.1
well-informed	0	24	0.0	6.0
literate	0	26	0.0	5.9
cultured	0	19	0.0	5.7
rational	0	[illegible]		
sensitive	8	134	2.0	5.9
thoughtful	14	121	5.0	7.7
affectionate	6	31	4.5	6.2
clever	54	30	5.8	4.8
funny	233	103	7.0	5.7
catchy	19	0	5.8	0.0

emotionally	0	111	0.0	8.6
artificially	0	52	0.0	7.9
fiercely	0	26	0.0	7.0
moderately	0	11	0.0	5.7
reasonably	0	54	0.0	5.7
culturally	0	12	0.0	5.5
[illegible]				
wonderfully	20	9	5.4	4.5
very	1707	596	5.6	4.0
too	476	76	5.4	2.8
damn	12	0	5.6	0.0
awfully	15	0	6.1	0.0
terribly	25	0	6.2	0.0

being	0	208	0.0	6.1
robot	0	77	0.0	6.1
agent	9	455	0.4	6.0
guess	0	35	0.0	5.5
conversation	0	88	0.0	5.1
creature	11	137	2.4	5.9
[illegible]	81	80	5.8	5.7
fellow	52	14	5.1	3.1
pass	67	9	5.2	2.2
wordplay	21	0	5.8	0.0
chap	47	0	5.9	0.0
twist	94	0	6.5	0.0
trick	166	0	6.7	0.0

Fig. 3. Sketch-diff table showing the difference in usage of the English adjectives *clever* and *intelligent*.

These core functions (inter-linked with a concordancer) have been subject to continuous development which has in the recent years focused on two major aspects: adaptation for parallel corpora (i.e. bilinguality) and adaptation to multi-word expressions (see [18,45]), so that the Sketch Engine now has both bilingual word sketches for parallel (or comparable) corpora and multi-word sketches showing collocations of arbitrary long headword-collocation combinations like *young man* or *utilize available resource*.

A substantial part of the Sketch Engine deals with corpus building for users. The Sketch Engine integrates dozens of third-party tools that allow researchers to quickly have their text converted into a searchable corpus, for many languages also automatically annotated with lemmas and part-of-speech tags. Underlying processing pipelines used for language-specific sentence segmentation, tokenization, character normalization and tagging or lemmatization represent years of efforts of bringing all of these tools into consistent shape – where the devil is in details which however have huge impact on the final usability of the data.

In this respect Adam's intentions were always to make it as easy as possible for the users to process their data so that they will not need to bother with technical details, but focus on their research. Even close to the end Adam was thinking of ways of facilitating Sketch Engine users. His last revision conducted several months before his death highlights following areas:

- Building Very Large Text Corpora from the Web
- Parallel and Distributed Processing of Very Large Corpora
- Corpus Heterogeneity and Homogeneity
- Corpus Evaluation
- Corpora and Language Teaching
- Language Change over Time
- Corpus Data Visualization
- Terminology Extraction

Lexical Computing Limited is committed to making these latest ideas come to fruition.

7 Lexicography

While collecting data for his DPhil thesis [19] (see Sect. 2), Adam canvassed a number of lexicographers for their views on his developing ideas. Could his theoretical model of how words convey meanings have applications in the practical world of dictionary-making? Thus began Adam's involvement with lexicography, which was to form a major component of his working life for the rest of his career, and which had a transformative impact on the field.

After a spell as resident computational linguist at Longman Dictionaries (1992–1995), Adam returned to academia. Working first with Roger Evans and then with David Tugwell at the University of Brighton, he implemented his ideas for word profiles as the WASP-bench, 'a lexicographer's workbench supporting state-of-the-art word sense disambiguation' [72] (see Sect. 4). The notion of the word sketch first appeared in the WASP-bench, and a prototype version was used in the compilation of the Macmillan English Dictionary [65], a new, from-scratch monolingual learner's dictionary of English. The technology was a huge success. For the publisher, it produced efficiency gains, facilitating faster entry-writing. For the lexicographers, it provided a rapid overview of the salient features of a word's behavior, not only enabling them to disambiguate word senses with greater confidence but also providing immediate access to corpus sentences which instantiated any grammatical relation of interest. And crucially, it made the end-product more systematic and less dependent on the skills and intuitions of lexicographers. The original goal of applying the new WASP-bench technology to entry-writing was to support an improved account of collocation. But the unforeseen consequence was the biggest change in lexicographic methodology since the corpus revolution of the early 1980s. From now on, the word sketch would be the lexicographer's first port of call, complementing and often replacing the use of concordances — a procedure which was becoming increasingly impractical as the corpora used for dictionary-making grew by orders of magnitude.

Lexicography is in a process of transition, as dictionaries migrate from traditional print platforms to electronic media. Most current on-line dictionaries are "horseless carriages" — print books transferred uncomfortably into a new medium — but models are emerging for new electronic artifacts which will show more clearly the relationship between word use and meaning in context, supported by massive corpus evidence. Adam foresaw this and, through his many collaborations with working lexicographers, he not only provided (print) dictionary-makers with powerful tools for lexical analysis, but helped to lay the foundations for new kinds of dictionaries.

During the early noughties, the primitive word sketches used in a dictionary project at the end of the 1990s morphed into the Sketch Engine (see Sect. 6) which added a super-fast concordancer and a distributional thesaurus to the rapidly-improving word sketch tool [48]. Further developments followed as Adam responded to requests from dictionary developers.

In 2007, a lexicographic project which required the collection of many thousands of new corpus example sentences led to the creation of the GDEX tool [40]. The initial goal was to expedite the task of finding appropriate examples, which would meet the needs of language learners, for specific collocational pairings. Traditionally, lexicographers would scan concordances until a suitable example revealed itself, but this is a time-consuming business. GDEX streamlined the process. Using a collection of heuristics (such as sentence length, the number of pronouns and other anaphors in the sentence, and the presence or absence of low-frequency words in the surrounding context), the program identified the "best" candidate examples and presented them to the lexicographer, who then made the final choice. Once again, a CL-based technology delivered efficiency gains (always popular with publishers) while making lexicographers' lives a little easier. There was (and still is) room for improvement in GDEX's performance, but gradually technologies like these are being refined, becoming more reliable and being adapted for different languages [51].

As new components like GDEX were incorporated into the Sketch Engine's generic version, the package as a whole became a de facto standard for the language-analysis stages of dictionary compilation in the English-speaking world. But this was just the beginning. Initially a monolingual resource based around corpora of English, the Sketch Engine gradually added to its inventory dozens, then hundreds, of new corpora for all the world's major languages and many less resourced languages too — greatly expanding its potential for dictionary-making worldwide. This led Adam to explore the possibilities of using the Sketch Engine's querying tools and multilingual corpora to develop tools for translators. He foresaw sooner than most that, of all dictionary products, the conventional bilingual dictionary would be the most vulnerable to the changes then gathering pace in information technology. Bilingual Word Sketches have thus been added to the mix [18].

Adam also took the view that the boundary between lexical and terminological data was unlikely to survive lexicography's incorporation into the general enterprise of Search. In recent years, he became interested in enhancing the Sketch Engine with resources designed to simplify and systematize the work of terminologists. The package already included Marco Baroni's WebBootCat tool for building corpora from data on the web [4]. WebBootCat is especially well adapted to creating corpora for specialized domains and, as a further enhancement, tools have been added for extracting keyword lists and, more recently, key terms (salient 2- or 3-word items characteristic of a domain). In combination, these resources allow a user to build a large and diverse corpus for a specific domain and then identify the terms of art in that field — all at minimal cost. A related, still experimental, resource is a software routine designed to identify, in the texts of a specialized corpus, those sentences where the writer effectively supplies a definition of a term, paving the way (when the technology is more mature) for a configuration of tools which could do most of the work of creating a special-domain dictionary.

Even experiments which didn't work out quite as planned shed valuable light on the language system and its workings. An attempt to provide computational support for the process of selecting a headword list for a new collocation dictionary was only partially successful. But the **collocationality** metric it spawned revealed how some words are more collocational than others — an insight which proved useful as that project unfolded [29].

Adam was almost unique in being equally at home in the NLP and lexicographic communities. A significant part of his life's work involved the application of NLP principles to the practical business of making dictionaries. His vision was for a new way of creating dictionaries in which most of the language analysis was done by machines (which would do the job more reliably than humans). This presupposed a radical shift in the respective roles of the lexicographer and the computer: where formerly the technology simply supported the corpus-analysis process, in the new model it would be more proactive, scouring vast corpus resources to identify a range of lexicographically-relevant facts, which would then be presented to the lexicographer. The lexicographer's role would then be to select, reject, edit and finalize [66]. A prototype version of this approach was the **Tickbox lexicography** [66] model used in the project which produced the DANTE lexical database [2].

Lexicographers would often approach Adam for a computational solution to a specific practical problem, and we have described several such cases here. Almost always, Adam's way of solving the problem brought additional, unforeseen benefits, and collectively these initiatives effected a transformation in the way dictionaries are compiled.

But there is much more. Even while writing his doctoral thesis, Adam perceived the fundamental problem with the way dictionaries accounted for word meanings. Traditional lexicographic practice rests on a view of words as autonomous bearers of meaning (or meanings), and according to this view, the meaning of a sentence is a selective concatenation of the meanings of the words in it. But a radically different understanding of how meanings are created (and understood) has been emerging since at least the 1970s. In this model, meaning is not an inherent property of the individual word, but is to a large degree dependent on context and co-text. As John Sinclair put it,

> *Many if not most meanings depend for their normal realization on the presence of more than one word* [71].

This changes everything — and opens up exciting opportunities for a new generation of dictionaries. Conventional dictionaries identify word senses, but without explaining the complex patterns of co-selection which activate each sense. What is in prospect now is an online inventory of phraseological norms and the meanings associated with them. A "dictionary" which mapped meanings onto the recurrent patterns of usage found in large corpora would in turn make it easier for machines to process natural language. Adam grasped all this at an early stage in his career, and the software he subsequently developed (from the WASP-Bench onwards) provides the tools we will need to realize these ambitious goals.

The cumulative effect of Adam's work with lexicographers over twenty-odd years was not only to reshape the way dictionaries are made, but to make possible the development of radically different lexical resources which will reveal — more accurately, more completely, and more systematically than ever before — how people create and understand meanings when they communicate.

8 Conclusions and Outlook

As this review article highlights, Adam made a huge scientific contribution, not just to the field of computational linguistics but in other areas of linguistics and in lexicography. Adam was a man of conviction. He was eager to hear and take on new ideas but his belief in looking carefully at the data was fundamental. He raised questions over common practice in WSD [23,25], the lack of due care and attention to replicability when obtaining training data [31] as well as assumptions in other areas [28,32]. Though our perspectives of his ideas and work will vary, there is no doubt that our field is the better for his scrutiny and that his ideas have been seminal in many areas.

Adam contributed a great deal more than just ideas and papers. He was responsible, or a catalyst, for the production of a substantial amount of software, evaluation protocols and data (both corpora and annotated data sets). He had a passion for events and loved bringing people together as evidenced by his huge network of collaborators, of which the authors of this article are just a very small part. He founded or co-founded many events including Senseval (now SemEval), the ACL's special interest group on Web as Corpus, and more recently the 'Helping Our Own' [7] exercise which has at its heart the idea of using computational linguistics to help non-native English speakers in their academic writing. This enterprise was typical of Adam's inclusivity.[13] He was exceptional in his enthusiasm for work on languages other than English and fully appreciated the need for data and algorithms for bringing human language technology to the masses of speakers of other languages, as well as enriching the world with access to information regardless of the language in which it was

[13] Other examples include his eagerness to encourage participants in evaluations such as Senseval, reminding people to focus on analysis rather than who came top [42] and in his company's aim of 'corpora for all'.

recorded. Adam was willing to go out on a limb for papers for which the standard computational linguistics reviewing response was 'Why didn't you do this in English for comparison?' or 'This is not original since it has already been done in English'. These rather common views mean that those working on other languages, and particularly less resourced languages, have a far higher bar for entry into computational linguistics conferences and Adam championed the idea of leveling this particular playing field.

Right to the very end, Adam thought about the future of language technology and particularly about possibilities for bringing cutting edge resources within the grasp of those with a need for them, but packaged in such a way as to make the technologies practical and straightforward to use. For specific details of the last ideas from Adam see the end of Sect. 6.

The loss of Adam is keenly felt. There are now conference prizes in his name, at eLex[14] and at the ACL *SEM conference. SIGLEX, of which he was president 2000–2004, is coordinating an edited volume of articles *Computational Lexical Semantics and Lexicography Essays: In honor of Adam Kilgarriff* to be published by Springer later this year. This CICLing 2016 conference is dedicated to Adam's memory in recognition for his great contributions to computational linguistics and the many years of service he gave on CICLing's small informal steering committee. There is no doubt that we will continue to benefit from Adam Kilgarriff's scientific heritage and that his ideas and the events, software, data and communities of collaborators that he introduced will continue to influence and enable research in all aspects of language technology in the years to come.

References

1. Atkins, S.: Tools for computer-aided corpus lexicography: the hector project. Acta Linguistica Hungarica **41**, 5–72 (1993)
2. Atkins, S., Rundell, M., Kilgarriff, A.: Database of ANalysed Texts of English (DANTE). In: Proceedings of Euralex (2010)
3. Banko, M., Brill, E.: Scaling to very very large corpora for natural language disambiguation. In: ACL, pp. 26–33 (2001)
4. Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P.: WebBootCat: a web tool for instant corpora. In: Proceedings of Euralex, Torino, Italy, pp. 123–132 (2006)
5. Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlỳ, P.: WebBootCaT: instant domain-specific corpora to support human translators. In: Proceedings of EAMT, pp. 247–252 (2006)
6. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI Lecture Notes. CSLI Publications, Stanford (2002). http://opac.inria.fr/record=b1098622
7. Dale, R., Kilgarriff, A.: Helping our own: text massaging for computational linguistics as a new shared task. In: Proceedings of the 6th International Natural Language Generation Conference, pp. 263–267. Association for Computational Linguistics (2010)

[14] See http://kilgarriff.co.uk/prize/.

8. Erjavec, T., Evans, R., Ide, N., Kilgarriff, A.: The concede model for lexical databases. In: Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 355–362. Athens, Greece (2000)
9. Evans, R., Gazdar, G.: DATR: a language for lexical knowledge representation. Comput. Linguist. **22**(2), 167–216 (1996). http://eprints.brighton.ac.uk/11552/
10. Gale, W., Church, K., Yarowsky, D.: One sense per discourse. In: Proceedings of the 4th DARPA Speech and Natural Language Workshop, pp. 233–237 (1992)
11. Gardner, S., Nesi, H.: A classification of genre families in university student writing. Appl. Linguist. **34**(1), 25–52 (2012). ams024
12. Gilquin, G., Granger, S., Paquot, M.: Learner corpora: the missing link in EAP pedagogy. J. Engl. Acad. Purp. **6**(4), 319–335 (2007)
13. Hanks, P.: Do word meanings exist? Comput. Humanit. **34**(1–2), 205–215 (2000). SENSEVAL Special Issue
14. Ide, N., Kilgarriff, A., Romary, L.: A formal model of dictionary structure and content. In: Heid, U., Evert, S., Lehmann, E., Rohrer, C. (eds.) Proceedings of the 9th EURALEX International Congress. Institut für Maschinelle Sprachverarbeitung, Stuttgart, Germany, pp. 113–126, August 2000
15. Ide, N., Véronis, J.: Encoding dictionaries. Comput. Humanit. **29**(2), 167–179 (1995). http://dx.doi.org/10.1007/BF01830710
16. Jakubíček, M., Rychlý, P., Kilgarriff, A., McCarthy, D.: Fast syntactic searching in very large corpora for many languages. In: PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Tokyo, pp. 741–747 (2010)
17. Kallas, J., Tuulik, M., Langemets, M.: The basic Estonian dictionary: the first monolingual L2 learner's dictionary of Estonian. In: Proceedings of the XVI Euralex Congress (2014)
18. Kilgarriff, A., Kovar, V., Frankenberg-Garcia, A.: Bilingual word sketches: three flavours. In: Electronic Lexicography in the 21st Century: Thinking outside the Paper (eLex 2013), pp. 17–19 (2013)
19. Kilgarriff, A.: Polysemy. Ph.D. thesis, University of Sussex (1992)
20. Kilgarriff, A.: Dictionary word-sense distinctions: an enquiry into their nature. Comput. Humanities **26**(1–2), 365–387 (1993)
21. Kilgarriff, A.: The hard parts of lexicography. Int. J. Lexicography **11**(1), 51–54 (1997)
22. Kilgarriff, A.: Putting frequencies in the dictionary. Int. J. Lexicography **10**(2), 135–155 (1997)
23. Kilgarriff, A.: What is word sense disambiguation good for? In: Proceedings of Natural Language Processing in the Pacific Rim, pp. 209–214 (1997)
24. Kilgarriff, A.: Gold standard datasets for evaluating word sense disambiguation programs. Comput. Speech Lang. **12**(3), 453–472 (1998)
25. Kilgarriff, A.: I don't believe in word senses. Comput. Humanit. **31**(2), 91–113 (1998). Reprinted in Practical Lexicography: a Reader. Fontenelle (ed.) Oxford University Press (2008). Also reprinted in Polysemy: Flexible patterns of meaning in language and mind Nerlich Todd, Herman and Clarke (eds.) Walter de Gruyter, pp. 361–392. And to be reprinted in Readings in the Lexicon Pustejovsky and Wilks (eds.) MIT Press
26. Kilgarriff, A.: SENSEVAL: an exercise in evaluating word sense disambiguation programs. In: Proceedings of LREC, Granada, pp. 581–588 (1998)
27. Kilgarriff, A.: Comparing corpora. Int. J. Corpus Linguist. **6**(1), 1–37 (2001)
28. Kilgarriff, A.: Language is never ever ever random. Corpus Linguist. Linguist. Theor. **1**(2), 263–276 (2005)

29. Kilgarriff, A.: Collocationality (and how to measure it). In: Proceedings of the 12th EURALEX International Congress, Torino, Italy, September 2006, pp. 997–1004 (2006)
30. Kilgarriff, A.: Word senses. In: Agirre, E., Edmonds, P. (eds.) Word Sense Disambiguation, Algorithms and Applications, pp. 29–46. Springer, Heidelberg (2006). https://doi.org/10.1007/978-1-4020-4809-8
31. Kilgarriff, A.: Googleology is bad science. Comput. Linguist. **33**(1), 147–151 (2007)
32. Kilgarriff, A.: Grammar is to meaning as the law is to good behaviour. Corpus Linguist. Linguist. Theor **3**(2), 195–197 (2007)
33. Kilgarriff, A.: Simple maths for keywords. In: Proceedings of Corpus Linguistics, Liverpool, UK (2009)
34. Kilgarriff, A.: Comparable corpora within and across languages, word frequency lists and the kelly project. In: Procedings of Workshop on Building and Using Comparable Corpora at LREC, Malta (2010)
35. Kilgarriff, A.: A detailed, accurate, extensive, available English lexical database. In: Proceedings of the NAACL HLT 2010 Demonstration Session, pp. 21–24. Association for Computational Linguistics, Los Angeles, June 2010. http://www.aclweb.org/anthology/N10-2006
36. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1**(1), 7–36 (2014). http://dx.doi.org/10.1007/s40607-014-0009-9
37. Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J.B., Khalil, S., Kokkinakis, S.J., Lew, R., Sharoff, S., Vadlapudi, R., Volodina, E.: Corpus-based vocabulary lists for language learners for nine languages. Lang. Resour. Eval. **48**(1), 121–163 (2014)
38. Kilgarriff, A., Evans, R., Koeling, R., Rundell, M., Tugwell, D.: WASPBENCH: a lexicographer's workbench supporting state-of-the-art word sense disambiguation. In: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics, EACL 2003, vol. 2, pp. 211–214. Association for Computational Linguistics, Stroudsburg (2003). https://doi.org/10.3115/1067737.1067787
39. Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on web as corpus. Comput. Linguist. **29**(3), 333–347 (2003)
40. Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P.: GDEX: automatically finding good dictionary examples in a corpus. In: Proceedings of the 13th EURALEX International Congress, Barcelona, Spain, July 2008, pp. 425–432 (2008)
41. Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., Suchomel, V.: Finding terms in corpora for many languages with the Sketch Engine. In: EACL 2014, p. 53 (2014)
42. Kilgarriff, A., Palmer, M.: Introduction to the special issue on SENSEVAL. Comput. Humanit. **34**(1–2), 1–13 (2000). SENSEVAL Special Issue
43. Kilgarriff, A., Palmer, M. (eds.): SENSEVAL98: Evaluating Word Sense Disambiguation Systems, pp. 1–2. Kluwer, Dordrecht (2000)
44. Kilgarriff, A., Rosenzweig, J.: Framework and results for English SENSEVAL. Comput. Humanit. **34**(1–2), 15–48 (2000). SENSEVAL Special Issue
45. Kilgarriff, A., Rychlý, P., Kovář, V., Baisa, V.: Finding multiwords of more than two words. In: Proceedings of EURALEX 2012 (2012)
46. Kilgarriff, A., Rychlý, P.: Semi-automatic dictionary drafting download. In: de Schryver, G.M. (ed.) A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks, Menha (2010)

47. Kilgarriff, A., Rychlỳ, P., Jakubicek, M., Kovár, V., Baisa, V., Kocincová, L.: Extrinsic corpus evaluation with a collocation dictionary task. In: LREC, pp. 545–552 (2014)
48. Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D.: The sketch engine. In: Proceedings of Euralex, Lorient, France, pp. 105–116 (2004). Reprinted in Patrick Hanks (ed.) (2007). Lexicology: Critical Concepts in Linguistics. Routledge, London
49. Kilgarriff, A., Tugwell, D.: WASP-Bench: an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation. In: Proceedings of the MT Summit VIII, Santiago de Compostela, Spain, pp. 187–190, September 2001
50. Kosem, I., Gantar, P., Krek, S.: Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In: Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference, Tallinn, Estonia, 17–19 October 2013, pp. 32–48 (2013)
51. Kosem, I., Husák, M., McCarthy, D.: GDEX for slovene. In: Proceedings of eLex2011, Bled, Slovenia (2011)
52. Krek, S., Abel, A., Tiberius, C.: ENeL Project: DWS/CQS Survey Analysis (2015). http://www.elexicography.eu/wp-content/uploads/2015/04/ENeL_WG3_Vienna_DWS_CQS_final_web.pdf
53. Leech, G.: 100 million words of English: the British national corpus (BNC). Lang. Res. **28**(1), 1–13 (1992)
54. Louw, B., Chateau, C.: Semantic prosody for the 21st century: are prosodies smoothed in academic contexts? A contextual prosodic theoretical perspective. In: Proceedings of the tenth JADT Conference on Statistical Analysis of Textual Data, pp. 754–764. Citeseer (2010)
55. Baroni, M., Chantree, F., Kilgarriff, A., Sharoff, S.: CleanEval: a competition for cleaning web pages. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 638–643 (2008)
56. Mautner, G.: Mining large corpora for social information: the case of elderly. Lang. Soc. **36**(01), 51–72 (2007)
57. McCarthy, D., Kilgarriff, A., Jakubíček, M., Reddy, S.: Semantic word sketches. In: 8th International Corpus Linguistics Conference (CL 2015) (2015)
58. McEnery, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh (1999)
59. Mihalcea, R., Chklovski, T., Kilgarriff, A.: The SENSEVAL-3 English lexical sample task. In: Mihalcea, R., Edmonds, P. (eds.) Proceedings SENSEVAL-3 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Barcelona, Spain, pp. 25–28 (2004)
60. Nastase, V., Sayyad-Shirabad, J., Sokolova, M., Szpakowicz, S.: Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21, no. 1, p. 781. AAAI Press/MIT Press, Menlo Park, Cambridge, London 1999 (2006)
61. O'Donovan, R., O'Neill, M.: A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In: Proceedings of the XIII EURALEX International Congress, Barcelona, 15–19 July 2008, pp. 571–579 (2008)
62. Peters, W., Kilgarriff, A.: Discovering semantic regularity in lexical resources. Int. J. Lexicography **13**(4), 287–312 (2000)
63. Pomikálek, J., Rychlỳ, P., Kilgarriff, A., et al.: Scaling to billion-plus word corpora. Adv. Comput. Linguist. **41**, 3–13 (2009)

64. Preiss, J., Yarowsky, D. (eds.): Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France (2001). sIGLEX Workshop Organized by Cotton, S., Edmonds, P., Kilgarriff, A., Palmer, M
65. Rundell, M.: Macmillan English Dictionary. Macmillan, Oxford (2002)
66. Rundell, M., Kilgarriff, A.: Automating the creation of dictionaries: where will it all end? In: Meunier, F. et al. (eds.) A Taste for Corpora. In Honour of Sylviane Granger, pp. 257–281. Benjamins, Amsterdam (2011)
67. Rychlý, P.: Korpusové manažery a jejich efektiví implementace. Ph.D. thesis, Masaryk University, Brno (únor 2000)
68. Rychlý, P.: Manatee/Bonito - a modular corpus manager. In: Proceedings of Recent Advances in Slavonic Natural Language Processing 2007. Masaryk University, Brno (2007)
69. Rychlý, P.: A lexicographer-friendly association score. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 6–9 (2008)
70. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In: Baroni, M., Bernardini, S. (eds.) WaCky! Working Papers on the Web as Corpus, Gedit, Bologna (2006)
71. Sinclair, J.: The lexical item. In: Weigand, E. (ed.) Contrastive Lexical Semantics. Benjamins, Amsterdam (1998)
72. Tugwell, D., Kilgarriff, A.: WASP-Bench: a lexicographic tool supporting word-sense disambiguation. In: Preiss, J., Yarowsky, D. (eds.) Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France (2001)
73. Tugwell, D., Kilgarriff, A.: Word sketch: extraction and display of significant collocations for lexicography. In: Proceedings of the ACL Workshop on Collocations, Toulouse, France, pp. 32–28 (2001)
74. Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., Saurí, R.: Classification of discourse coherence relations: an exploratory study using multiple knowledge sources. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, SigDIAL 2006, pp. 117–125, Association for Computational Linguistics, Stroudsburg (2006). http://dl.acm.org/citation.cfm?id=1654595.1654618
75. Yarowsky, D.: One sense per collocation. In: Proceedings of the ARPA Workshop on Human Language Technology, pp. 266–271. Morgan Kaufman (1993)
76. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189–196 (1995)