Adam Kilgarriff: Structured bibliography

Links

- 1. Preamble
- 2. Word Senses
  - 2.1 SENSEVAL/SEMEVAL
  - 2.2 Software responses
  - 2.3 Sketch Engine
- 3. Corpora
  - 3.1 Corpus building and web corpora
- 4. Supplementaries
  - 4.1 Dictionary-making
  - 4.2 Language-specific work
  - 4.3 Language learning and teaching
  - 4.4 Formal inheritance lexicons
  - 4.5 Use and abuse of statistics
- 5. Leftovers

1. Preamble

I have wondered about writing a book, modelled on Patrick Hanks's *Lexical Analysis*, in which he presents his 'Theory of Norms and Exploitations', as developed over his academic career. The book's chapters are lightly-edited versions of published papers from the last three decades, assembled in logical rather than chronological order and with some additional text to trace the main arguments through.

This is a first step towards a book like that. I have not edited the papers at all, and the links in this page are to the published versions of papers as found in my full bibliography (2009 and earlier, *here*; from 2010, *here*). But they are structured according to the distinct threads I have been following, with a little linking text (and with minor, work-in-progress and repeat items, or ones where I had a small role, omitted.)

My driving research questions have been what is a word sense (earlier in my career) and how can we get scientific about the different types of language that there are (more recently). My work on word senses moved from analytic to synthetic: from "what the problems are" to the SENSEVAL (later SEMEVAL) initiative and software responses. My work on different types of language is (needless to say, probably, to readers who have got this far) all based on corpora, with each text type represented by a corpus of it. A language (like English) comprises core and sublanguages: how can we turn that into a scientific statement, in terms of the statistical shape of corpora of different kinds?

The supplementary areas where I may have something to offer include dictionary-making, word sketches for particular languages (co-authored with an expert in that language), the role of corpora in language learning and teaching, formal inheritance lexicons, and the use and abuse of statistics.

In the below, I first give what I see as my main contributions to the two main research questions, and then move on to the supplementaries.
2. What is a word sense

This was the topic of my thesis, titled simply "Polysemy" (1992). While that would be one place to start reading, it might be better to start with a journal-article version of the main argument, "I don't believe in word senses" (1997) or a piece written explicitly to present the philosophical banana skins that come with 'word meaning': the "Word Senses" chapter of Agirre and Edmonds's book on Word Sense Disambiguation: Algorithms and Application (2006). My empirical work on dictionaries is described in "Dictionary Word Sense Distinctions: An enquiry into their nature" (1993).

From the point of view of someone who likes to be able to measure things, two key features of word senses are: the more text you look at, the more of them you find; and, their distribution tends to be highly skewed, with the commonest sense accounting for a large part of the data. I built a mathematical model to explore these assertions: see "How dominant is the commonest sense of a word?" (2004).

One thing that an academic combatant must do is confront the opposing theories: I don't do much of it, but I was sufficiently engaged and excited by the ideas of Pustejovsky's Generative Lexicon to do just that in "Generative lexicon meets corpus data: the case of non-standard word uses."

Two short and focussed explorations of word senses in the context of mid-1990s NLP are "Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction" (1997) and "What is Word Sense Disambiguation Good For?" (also 1997). I think the former is still salient now; on the latter, the topic is explored in detail in the Agirre and Edmonds WSD book (from 2006, mentioned above), and Machine Translation has moved on so far and fast that my mid-1990s optimism for the role of WSD there looks quite displaced.

2.1 SENSEVAL (/SEMEVAL)

Evaluation was a new and exciting topic in NLP in the 1990s. I had the good fortune to be in a good position to bring the 'competitive evaluation' model to Word Sense Disambiguation (WSD). This is the model where someone sets up a task and then welcomes anyone to enter their system into the competition to see how well it can do. I set up the task (much aided and supported by Martha Palmer) and we had lots of participants, many of whom came to the workshop at Herstmonceux Castle where we announced and discussed results. A brief account of the exercise is "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs" (1998). The full version is a journal special issue, see "Introduction to the Special Issue on SENSEVAL" (with Martha Palmer) and journal article "English Framework and Results" (with Joseph Rosenzweig, both 2000). Three years later, in 2001, we had the next SENSEVAL, which I co-ordinated with Phil Edmonds and which gave rise to another journal special issue: "Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems" (2002). The exercise continues to thrive, now under the name of "SEMEVAL."

2.2 Software Responses

I had done the analysis of word senses: the next step was a constructive response. How could an understanding of the nature of word senses give rise to good software for discovering and managing them?

The WASPS project developed and evaluated one model "An evaluation of a lexicographer's workbench: Building lexicons for machine translation" (2003, with Rob Koeling, David Tugwell, Roger Evans). It proved very hard to make it work well. I tried again a number of years later, this time with Pavel Rychly: "Semi-automatic Dictionary Drafting" (2010). As the acronym shows, it still proved too hard a task.
While semi-automating the task was too ambitious, it turned out we could do the first part of the task - identifying grammatical patterns and collocations for each word automatically from corpora - very well. We called the summaries word sketches and they were first used to support the lexicography, over the period 1999-2002, of a brand new dictionary, the Macmillan English Dictionary for Advanced Learners (edited by Michael Rundell, 2002). The process is described in Lexical profiling software and its lexicographic applications - a case study and was presented at the EURALEX lexicography conference, in Copenhagen in 2002.

2.3 The Sketch Engine

The work was enthusiastically received, and afterwards a number of people came up to me and asked "can I have them for my language please". My immediate response was "sorry, no, we don't have a corpus like the British National Corpus (which all the English work has been based on) for your language" - but after a while that began to feel rather feeble. So I applied myself to a better answer.

The best way to think of word sketches was as an additional feature of a corpus query system - so I needed to find a corpus query system to which word sketches could be added, and a computer scientist to do the adding. I was very lucky to find Pavel Rychly, who had recently developed a corpus query system, and who was keen to co-develop with me. Pavel extended and developed his system to incorporate word sketches, and we launched the new system, the Sketch Engine as a product of my company, Lexical Computing Ltd., at EURALEX in Lorient, France, in 2004.

Pavel and I have been working together since, with a growing team who have largely been his PhD students. We evaluate word sketches in A Quantitative Evaluation of Word Sketches? (2010, with Vojtech Kovar, Simon Krek, Irena Srdanovic, Carole Tiberius). We review the Sketch Engine, what it offers, and what has happened in the decade since its launch, in The Sketch Engine: Ten Years On.

We have now developed word sketches for many languages: associated publications, always with a co-author who is a language expert, are listed below.

3. Corpora

If accounts of language are to be based on corpora, then they will only be as good as the corpora they are based on. If linguistics is to be an objective science, then it should be based on samples of data: for good science we need to take the sampling seriously. The sampling is the corpus development.

In the mid 1990s I investigated the various statistics that were used with corpora: Which words are particularly characteristic of a text? A survey of statistical approaches and also Why chi-square doesn't work, and an improved LOB-Brown comparison (both 1996). It became clearer and clearer to me that some scheme of measuring distances between text types, as represented by corpora, was needed. Without one, the foundations of corpus methods for linguistics would remain shaky. I make a proposal, and bring together my 1990s work on the topic, in my main theoretical contribution: Comparing Corpora (2001).

One big challenge for corpus-users is Getting to know your corpus (2012) so this paper provides some advice.

3.1 Corpus Building and Web Corpora

A full account of an early corpus-building experience is "Efficient corpus development for lexicography: building the New Corpus for Ireland" (2006, with Michael Rundell and Elaine Úí Dhonnchadha). The stages are: design, collection, encoding.

The BNC design has been used as a model for many other corpora, even though it was from 1990 and the world has moved on: see "BNC Design Model Past its Sell-by" (2007, with Sue Atkins and Michael Rundell).

From the late 1990s, it was apparent that the web transformed opportunities for corpus building. The web was itself, after a manner, a corpus. The name of a short 2001 piece I wrote, "Web as Corpus," stuck as the name for a thread of research and a "Special Interest Group and workshop series on how to use the web (or parts of it) as a corpus. With Greg Grefenstette, I edited a Special Issue of *Computational Linguistics* on "Web as Corpus" and wrote the "Introduction" (2003). The Special Interest Group set up a competitive evaluation on 'cleaning' web pages; "CleanEval: a competition for cleaning web pages" (2008, with Marco Baroni, Francis Chantree, Kilgarriff and Serge Sharoff). We developed methods for creating small, medium and large corpora from the web:

- small and specialised: "WebBootCaT: a web tool for instant corpora" (2006, with Marco Baroni, Jan Pomikalek, Pavel Rychly)
- largish, for smaller languages: "A Corpus Factory for many languages" (2010, with Siva Reddy, Jan Pomikalek, Avinesh PVS)
- large, for big languages
  - of purely historical interest
    - "Linguistic Search Engine" (2003)
    - "Large linguistically-processed Web corpora for multiple languages" (2006, with Marco Baroni)

In a squib called "Googleology is Bad Science" I explored the use of Google (and other search engines) in corpus building. Corpus building is similar to what Google does and it is instructive to see how they relate.

4. Supplementaries

4.1 Dictionary-making

I have worked closely with dictionary publishers, with two years (1993-95) employed at Longman and extended periods of consultancy for OUP and Macmillan. My work has been of interest to them, and their work has been of interest to me.

One question about lexicography which puzzled me is "which parts of it are difficult?" I conducted a survey, amongst professionals, and reported results in a short piece: "The hard parts of lexicography" (1997).

While at Longman, I was given the task of adding frequency-markers into the third edition of LDOCE (Longman Dictionary of Contemporary English). I then wrote up the exercise, including discussion of all the stumbling blocks and choice points, in "Putting frequencies in the dictionary" (1997).

One recurring challenge for lexicographers and dictionary publishers is "how to choose good examples?" (and the underlying question "what makes a good example?"). I and colleagues worked with Macmillan to, first, develop,
and then, apply a semi-automatic response. The project delivered a large set of extra example sentences to an online version of Macmillan English Dictionary, and is written up as ?GDEX: Automatically finding good dictionary examples in a corpus (2008, with Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý).

As dictionaries go electronic, there is no longer a space constraint. Users like examples, so there is no longer a reason not to give them - provided we can provide good ones, cheaply. Correspondingly, there has been a high level of interest in GDEX, with follow-up papers for Estonian and Slovene, listed under the language-specific section below, and it forming a thread of the ?eNeL EU research network. It has been integrated into the Sketch Engine, and SKELL (see below), and has given rise to a proposal for a further level of semi-automation of lexicography: ?Ticbox Lexicography(2010, with Vojtech Kovar and Pavel Rychly).


What should general language dictionaries do about the ineluctably composite nature of language, whereby any language (Dutch, or English, or Chinese) comprises a range of genres and sublanguages. As discussed above, the broad question is at the core of my research. As advisor on ?A Frequency Dictionary for Dutch, I had the opportunity to put some ideas into practice, see ?Genre in a frequency dictionary (2013, with Carole Tiberius).

Until recently, dictionary publishers have only looked at word sketches as a way of supporting dictionary creation. A next stage is to include them in the online dictionary, as shown to the user. This may also support Search Engine Optimisation: the challenge of getting your web page to the top of Google rankings. We ran an experiment, working with OUP, and reported on it in ?Augmenting Online Dictionary Entries with Corpus Data for Search Engine Optimisation (2013, with Holger Hvelplund, Vincent Lannoy, Patrick White).

As already noted, sometimes one must respond to the opposition. Here I address the 'Aarhus School' of lexicographic function theory: Review of: Pedro A. Fuertes-Olivera and Henning Bergenholtz (eds.) e-Lexicography: The Internet, Digital Initiatives and Lexicography

4.2. Language-specific work: word sketches, corpus building, lexicography

- Arabic

- Chinese
  - ?Chinese Sketch Engine and the Extraction of Grammatical Collocations (2005, with Chu-Ren Huang, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, Keh-Jiann Chen)

- Czech
  - ?Czech Word Sketch Relations with full syntax parser (2009, with Ales Horak, Pavel Rychly)

- Dutch
  - ?The Sketch Engine for Dutch with the ANW corpus (2009, with Carole Tiberius)

- English
  - ?Database of ANalysed Texts of English (DANTE): the NEID database project (2010, with Sue Atkins and Michael Rundell)
  - ?The Oxford Children's Corpus: Using a Children's Corpus in Lexicography (2012, with Kate Wild, David Tugwell)

4.1 Dictionary-making
4.3 Language learning and teaching

My career owes a lot to English Language Teaching. Without a few crumbs from its large and well-laden table, the company might not have been a viable proposition. Most of the interaction has been mediated via ELT dictionaries: the dictionaries have been big earners for their publishers, so the publishers have been keen to keep their products top-of-the-range, and the Sketch Engine helped there. But that is indirect use, where learners’ use of the corpus is mediated by the dictionary. What about direct use, where learners look at the corpus themselves? This has been the core concern of the Teaching and Language Corpora (TALC) community for three decades.

Language teachers have been using the Sketch Engine in the classroom, and I have been asked to talk about it at a number of venues. I first assembled my thoughts in ?Corpora in the classroom without scaring the students (2009). I collaborated with Simon Smith on ?Automatic Cloze Generation for English Proficiency Testing (2009, also with Scott Sommers, Gong Wen-liang, Wu Guang-zhong) and ?Making better wordlists for ELT: Harvesting vocabulary lists from the web using WebBootCat, (2008, also with Scott Sommers).
In Learning Chinese with the Sketch Engine (2014, with Simon, Nicole Keng and Wei Bo) we make the case for applying what has been learnt about using corpora in ELT, to CLT (Chinese Language Teaching).

James Thomas, meanwhile, has been preparing a textbook: Discovering English with Sketch Engine.

The dominant feedback we have had on the Sketch Engine, from everyone except professional lexicographers and dedicated corpus linguists, is **too many buttons**. The shout has come loudest from the ELT community. In response to that we have produced a slimmed-down version: **SKELL** (Sketch Engine for Language Learning).

James, Simon and I review how corpora are being used in language teaching, and provide an introduction to SKELL in Corpora and Language Learning with the Sketch Engine and SKELL (2015, also with Fredrik Markowitz).

One straightforward way for corpora to support language learning is for corpora to provide word lists, of the common words in a language, to be used in syllabus design, creation and selection of reading materials for learners, and language testing, as well as for deciding which words to put in dictionaries. In the popular ‘word cards’ approach to vocabulary teaching, it should be the common words of each language on the cards. In the EU project Kelly, we created ?Corpus-Based Vocabulary lists for Language Learners for Nine Languages (2013, with Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, Elena Volodina) with this agenda.

### 4.4 Formal inheritance lexicons

When I started my PhD, my supervisors, Gerald Gazdar and Roger Evans, were both working on a knowledge representation language called **DATR**, designed principally for lexicons and based on default inheritance. I developed DATR accounts of verb alternations (?Inheriting Verb Alternations 1993) and noun polysemy (?Inheriting Polysemy, also 1993).

While I went on to get much more engaged in corpus methods than formal ones, the basic truth that we humans organise our knowledge hierarchically continues to tease me. I still dream of connecting hierarchical knowledge to the statistical models that corpora can provide.

### 4.5 Use and abuse of statistics

Language is never, ever, ever, random. This bald fact means that it is inappropriate to use statistical hypothesis-testing in some places where they often have been used, as the paper explains. One thing that corpus users often want to do is to find the keywords in one text (or corpus) versus another. A simple method is presented in ?Simple maths for keywords.

### 5. Leftovers

- **How Many Words Are There?** (2014)
- **Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine**