Sketching words

Adam Kilgarriff and David Tugwell

(ITRI, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK)

Abstract

This paper introduces the Word Sketch: a summary of a word's grammatical and collocational behaviour produced automatically, from a large corpus, for a lexicographer.[1] The Word Sketch improves on standard collocation lists by using a grammar and a parser to find collocates in specific grammatical relations, and then producing one list of subjects, another of objects, etc, rather than a single grammatically blind list. The Word Sketches have been used in a a large dictionary project and have received positive reviews.

1. Four Ages of Corpus Lexicography

The first age of corpus lexicography was pre-computer. Samuel Johnson collected citations; the Oxford English Dictionary had a corpus of five million index cards, each with a citation on. Sue Atkins was not involved in the inauguration of that era.

The second age dawned with the COBUILD project, circa 1980, brainchild of Sue and her brother, John Sinclair. Computers could be used to store text, and to produce concordances. Lexicographers would thereby be able to view the evidence of how a word was used without the arbitrary filter of who thought what was an interesting example of a word.

As all readers are doubtless aware, the use of computerised corpora has transformed lexicography. Any forward-looking dictionary project uses one. It either negotiates to use a corpus that already exists or creates one afresh. (If it creates one afresh, the model, or rather object of desire that the dictionary project would replicate for their own language if only they had the resources, is usually the British National Corpus[2] (BNC), another Atkins baby).

Following Sue's move to Oxford University Press in 1989, OUP too, were addressing the challenge of using a corpus. In her search for ideas that would help her make better use of the corpus to make better dictionaries, Sue sought collaborators and made friends in the computational world, at IBM, AT&T and Digital. The liaison with Digital developed into the HECTOR project, in which

[1] This work was supported by the UK EPSRC under grant M54971 (WASPS).

[2] http://info.ox.ac.uk/bnc

Sue (with Patrick Hanks) organised an astonishing range of resources for their lexicographers. (Sue's demands for computational support have never been modest. As she puts it "after three weeks of trying to explain to me it's impossible, the programmers realise the only way they'll get any peace is to just write the system so then they get on with it.") With a 20 million word corpus from the American Publishing House for the Blind, parsed using Don Hindle's Fidditch parser (itself a major innovation in computational linguistics, arguably the first wide-coverage parser), a corpus access system that in due course turned into the Altavista search engine, and a setup of no less than three co-ordinated computer screens for lexicographers to view, HECTOR was a visionary project. It did not immediately result in a dictionary or many publications, but it did chart the way ahead for corpus lexicography in general and developments such as Word Sketches in particular.

1.1. Statistical summaries

Where there are fifty instances for a word, the lexicographer can read them all. Where there are five hundred, they could, but the project timetable will rapidly start to slip. Where there are five thousand, it is definitely no longer feasible. Corpus query languages with sophisticated querying such as Xkwic [Schulze and Christ1994] help, but there is still too much data to view.

The third age of corpus lexicography was a response to this problem. The data needed summarising.

The answer, arising out of the collaboration with AT&T, was a statistical summary. The task is to look at the other words in the neighbourhood of the word of interest, its 'collocates', and to identify those that occur with interestingly high frequency in that neighbourhood. The statistic can be used to sort the collocates, and if the statistic (and the corpus) are good ones, the collocates that the lexicographer should consider mentioning percolate to the top.

Ken Church and Patrick Hanks proposed two statistics, pointwise Mutual Information and the t-score (which can be used both for identifying collocates, and for identifying how the collocates of two words of similar meaning differ). The paper describing the work [Church and Hanks1989] inaugurated a subfield of computational linguistics, "collocation statistics", and contributed to the decisive arrival of corpora in the field of computational linguistics.

Since Church and Hanks's proposals a series of papers have proposed alternative statistics [Dunning1993, Pedersen1996] (see [Kilgarriff1996] for a critical review), and evaluated different statistics [Evert and Krenn2001].

Now, any dictionary projects with access to a corpus provides statistical summaries to lexicographers. These contain many nuggets of information, but are not used as widely as they might be. From a lexicographical perspective, they have three failings. First, the statistics..They have not been ideal, with too many low frequency words occurring at the tops of the lists. Second, noise. Alongside the lexicographically interesting collocates are assorted uninteresting ones: words that do happen to occur in the neighbourhood of the nodeword, but do not stand in a linguistically interesting relation to it. Third, the neighbourhood. When searching for some types of collocates such as subjects for verbs in English, we wish to look for collocates preceding the nodeword, but it is not clear whether we should look at a window of one, three or five words prior to the nodeword, and possibly we should look at all of these, and in any case we are likely to find assorted adverbs, subjects of passives and other items mixed in with the subjects. It would be far more satisfactory to explicitly produce a collocate list for subjects, another for objects, and so forth (which would also eliminate most noise), as has been proposed by [Hindle1990] and [Tapanainen and Järvinen1998]. The Word Sketches are a large scale implementation of such improved collocate-lists for practical lexicography.

2. The Word Sketch Workbench

In this section we describe how the Word Sketches are produced, and how the lexicographers interacts with the system that builds them.

The workbench is implemented in perl. It uses cgi-scripts and a browser for user interaction, so is designed for client-server use, where the client may be local or remote and needs no software loaded onto it other than Netscape, Internet Explorer, or some other web browser.[3]

2.1. Grammatical relations database

The central resource is a collection of all grammatical relations holding between words in the corpus. The workbench is currently based on the British National Corpus (BNC): 100 million words of contemporary British English, of a wide range of genres. In its published form, the BNC is part-of-speech-tagged, by Lancaster's CLAWS tagger. These tags were used. The BNC was lemmatised, by the `morph` program [Minnen et al.2000]. Using a shallow parser implemented as a regular-expression matcher over part-of-speech tags, we processed the whole corpus to find quintuples of the form:

{Rel, Word1, Word2, Prep, Pos}

where Rel is a relation, Word1 is the lemma of the word for which Rel holds, Word2 is the lemma of the other open-class word involved, Prep is the

[3] A demo is available at http://wasps.itri.bton.ac.uk

preposition or particle involved and Pos is the position of Word1 in the corpus.[4] Relations may have null values for Word2 and Prep. The database currently contains approximately 70 million quintuples.

The current inventory of relations is shown in Table 1. These fall into the following classes:

- Nine unary relations (ie. with Word2 and Prep null). Three of these are exclusively for nouns (bare-noun, possessed and plural), two for verbs (passive and reflexive), while the remaining four complementation patterns are available for any word class. Unary relations may be seen to be of limited use by themselves for lexicography, but they will come into play where patterns are combined, as outlined in section 2.5.
- Seven binary relations with Prep null. Two of these are exclusively for verbs (object and adjectival complement), one for verbs and adjectives (subject), two for nouns (noun modifier and predicate), and two for all word classes (modifier and "and-or"). In addition, for six of these binary relations we also explicittltly represent the inverse relation, ie. subject-of etc, found by taking Word2 as the head word instead of Word1. The conjunction relation and-or is considered symmetrical so does not give rise to a separate inverse relation.
- Two binary relations with Word2 null. The preposition here is either a particle or introduces a gerundive phrase, and the relations may apply to any word class.
- One trinary relation, prepositional complement or modifier, which applies to all word classes. Taking Word2 as primary again, the inverse relation is also explicitly represented and may be glossed as "Word1 is head of the complement of a PP modifying Word2". The inverse relation is only applicable to nouns.

The number of relations, including inverse relations, is twenty-six.

It is also the case that the same instance may have more than one relation of the same kind, as in "banks, mounds and ditches" where *bank* has two **and-or** relations, one with *mound* and one with *ditch*, or "he saw the bank she had climbed" where *bank* has an **object-of** relation to both *see* and *climb*.

These relations provide a flexible resource which is used as the basis of the computations for the Word Sketch. It is similar to the database of triples used in [Lin1998] for thesaurus generation. Keeping the position numbers of examples allows us to find associations between relations, as outlined in section 2.5, and to display the actual context of use in the corpus.

[4] We store the corpus in the representation formalism developed at IMS Stuttgart [Schultze and Christ 1994]

relation	Example
bare-noun	the angle of **bank**^{1}
possessed	my **bank**^{1}
plural	the **banks**^{1}
passive	was **seen**^{1}
reflexive	**see**^{1} herself
ing-comp	**love**^{1} eating fish
finite-comp	**know**^{1} he came
inf-comp	**decision**^{1} to eat fish
wh-comp	**know**^{1} why he came
subject	the **bank**^{2} **refused**^{1}
object	**climb**^{1} the **bank**^{2}
adj-comp	**grow**^{1} **certain**^{2}
noun-modifier	**merchant**^{2} **bank**^{1}
modifier	a **big**^{2} **bank**^{1}
and-or	**banks**^{1} and **mounds**^{2}
predicate	**banks**^{1} are **barriers**^{2}
particle	**grow**^{1} **up**^{p}
Prep+gerund	**tired**^{1} **of**^{p} eating fish
PP-comp/mod	**banks**^{1} **of**^{p} the **river**^{2}

Table 1: Grammatical Relations

The relations contain a substantial number of errors, originating from POS-tagging errors in the BNC, limitations of the pattern-matching grammar or attachment ambiguities. Indeed no attempt is made to resolve the latter: "see the man with a telescope" will give rise to both {PP,*see*,*telescope*,*with*} and {PP,*man*,*telescope*,*with*}. However, as the system finds high-salience patterns, given enough data, the noise does not present great problems for the task in hand.

2.2. *Word Sketch Display*

When a lexicographer embarks on composing the lexical entry for a word, they enter the word (and word class) at a prompt. At present, word classes covered are noun, verb and adjective. Using the grammatical relations database, the system then composes a **Word Sketch** for the word. This is a page of data such as Table 2, which shows, for the word in question (Word1), ordered lists of high-salience grammatical relations, relation-Word2 pairs, and relation-Word2-Prep triples for the word. These are listed for each relation in order of salience, with the count of corpus instances. The actual corpus examples illustrating each pattern are available by mouse-click. Producing a Word Sketch for a medium-high frequency word currently takes around ten seconds

2.3. Calculating Salience

Salience is estimated as the product of Mutual Information ***I*** [Church and Hanks1989] and log frequency. ***I*** for a word ***W1*** in a grammatical relation $\mathbf{R}$[5] is calculated as

$$I(W1;R) = log\left(\left\|\frac{\|*,*,*\|\times\|W1,R,*\|}{\|W1,*,*\|\times\|*,R,*\|}\right\|\right)$$

The notation here is adopted from [Lin1998] (who also spells out the derivation from the definition of *I*). $\|W1,R,W2\|$ denotes the frequency count of the triple $\{W1,R,W2\}$[6] in the grammatical relations database. Where $W1$, R or $W2$ is the wild card $(*)$, the frequency is of all the dependency triples that match the remainder of the pattern.

Again following Lin, we calculate *I* for triples relative to the frequency of R:

$$I(W1,R,W2) = log\left(\left\|\frac{\|*,R,*\|\times\|W1,R,W2\|}{\|W1,R,*\|\times\|*,R,W2\|}\right\|\right)$$

The Word Sketches are presented to the user as a list of relations, with items in each list ordered according to salience. Thus it is not problematic that all calculations of *I* for triples are relative to $\|*,R,*\|$, the overall frequency of the relation. Arguably, *I*(*W*1,*R*,*W*2)should not be defined to be relative in this way.

Our experience of working lexicographers' use of collocate lists sorted by values of the Mutual Information or log-likelihood statistic shows that, for lexicographic purposes, this over-emphasises low frequency items. This is also the experience of lexicography projects at CUP, Collins, Longman and elsewhere. Multiplying by log frequency is an appropriate adjustment bringing words that are of greatest lexicographic relevance to the head of the collocate list.

2.4. Using Word Sketches

Table 2 shows a Word Sketch for the noun *bank*. It is slightly abbreviated due to the constraints of space, but is otherwise not modified or edited in any way. The total number of patterns shown for the word is set by the user according to needs.

[5] {Grammatical-relation, preposition} pairs are currently treated as atomic relations for purposes of caculating MI.

[6] The trinary relation, PP-comp/mod, is reduced to a set of binary relations by having a separate relation for each preposition, viz, *PP-com/mod:to*.

subject-of	num	sal
lend	95	21.2
issue	60	11.8
charge	29	9.5
operate	45	8.9
step	15	7.7
deposit	10	7.6
borrow	12	7.6
eavesdrop	4	7.5
finance	13	7.2
underwrite	6	7.2
account	19	7.1
wish	26	7.1

object-of	num	sal
burst	27	16.4
rob	31	15.3
overflow	7	10.2
line	13	8.4
privatize	6	7.9
defraud	5	6.6
climb	12	5.9
break	32	5.5
oblige	7	5.2
sue	6	4.7
instruct	6	4.5
owe	9	4.3

modifier	num	sal
central	755	25.5
Swiss	87	18.7
commercial	231	18.6
grassy	42	18.5
royal	336	18.2
far	93	15.6
steep	50	14.4
issuing	23	14.0
confirming	13	13.8
correspondent	15	11.9
state-owned	18	11.1
eligible	16	11.1

inv-PP	num	sal
governor of	108	26.2
balance at	25	20.2
borrow from	42	19.1
account with	30	18.4
account at	26	18.1
customer of	18	14.9
bank to	13	13.2
debt to	18	13.1
deposit at	9	12.3
pay into	14	12.0
branch of	34	11.2
loan by	6	10.7
situate on	14	10.6
subsidiary of	12	9.9
tree on	11	9.8
syndicate of	6	9.8
cash from	9	9.7
owe to	12	9.6

modifies	num	sal
holiday	404	32.6
account	503	32.0
loan	108	27.5
lending	68	26.1
deposit	147	25.8
manager	319	22.2
Holidays	32	21.6
clerk	73	21.4
balance	93	21.3
overdraft	23	20.3
robber	28	19.9
robbery	33	19.4
governor	41	17.0
debt	35	15.3
borrowing	21	15.2
note	65	15.2
credit	51	15.0
vault	19	13.9

noun-mod	num	sal
merchant	213	29.4
clearing	127	27.0
river	217	25.4
creditor	52	22.8
Tony	57	21.4
AIB	23	20.9
savings	61	19.8
Whinney	17	19.7
piggy	21	18.5
bottle	34	17.4
investment	121	17.0
August	39	16.8
canal	36	16.0
memory	57	16.0
Jeff	14	15.9
south	58	14.8
correspondent	13	14.5
shingle	16	14.4

and-or	num	sal
society	287	24.6
bank	107	17.7
institution	82	16.0
Bank	35	14.4
Lloyds	11	14.1
bundesbank	10	13.6
company	108	13.6
currency	26	13.5
issuing	7	13.0
Barclays	9	12.7
ditch	14	12.2
broker	15	11.3
lender	13	11.0
stockbroker	10	10.7

PP of	Num	sal
England	988	37.5
Scotland	242	26.9
river	111	22.1
Thames	41	20.1
credit	58	17.7
Severn	15	16.8
Japan	38	16.8
Ireland	56	16.0
Crete	14	15.3
stream	25	14.8
Nile	14	13.7
Montreal	11	13.4
cloud	22	12.7
River	12	12.3

PP for	num	sal
settlement	19	12.8
reconstruction	10	11.1

predicate	num	sal
bank	5	7.5
institution	4	5.6

predicate-of	num	sal
bank	5	6.0
country	6	4.3

plural	6760	2.3
bare noun	442	-9.0
possessed	639	-5.5

Table 2: Word Sketch for *bank* (n), BNC frequency = 20,968

Table 2 reveals how the different word senses for the word can be brought out as they tend to occur with particular significant patterns. For example as object of *burst* we have the RIVER BANK sense of the word, while the object of *rob* is the FINANCIAL INSTITUTION sense. Fixed idioms, such as *bank holiday*, are also revealed. While these are obvious senses, the Word Sketch also reveals less obvious ones, such as those in the collocations *bottle bank*, *bank of cloud*, *memory bank* etc. This should then be enough to serve as the basis for drawing up the lexical entry for the dictionary.

The 'number of examples' column in the Word Sketch contains a hyperlink to a collocation window. Clicking on the link brings up the actual examples from the BNC which contain the pattern in question, thus allowing the original corpus data to be examined. At the same time, examples may be pasted into lexical entries.

2.5. Combining Patterns

Consider the reduced Word Sketch for the verb *fall* given in Table 3.[7] A salient PP-pattern such as **into hand** may not be immediately recognisable as it is just composed of the preposition and the head of its complement noun phrase. A look at the corpus examples reveals that these are practically all of the form "into the hands of..." or "into someone's hands". Using the data we already have available we are in a position to calculate more fine-grained patterns revealing this by checking the other grammatical relations that hold for either Word1 or Word2 in the relation. Such a check will reveal that for Word2 in this pattern, other relations that hold in an overwhelming number of cases are **plural** and *possessed*. The pattern may be better presented then as **into sb's hands**.

Similarly for **by wayside**, Word2 will be exclusively definite and singular[8], allowing the pattern to be presented as **by the wayside**. Again a particular idiom of **into the trap of V-ing** may be identified by similar means.

The extra calculation involved in this refinement of collocational patterns is small, since it is confined to that small number of patterns which are already found to be of high salience. The fact that patterns in the database are explicitly marked with an instance number for Word1 marking its position in the corpus

[7] This sketch also illustrates some of the problems introduced by incorrect tagging in the original corpus: the collocation "fall short" appears in the patterns verb + object and verb + adverbial modifier, as well as the correct verb + adjectival complement. Indeed, all the verb + object patterns do not involve genuine objects, but are nevertheless useful to the lexicographer as being significant collocations.

[8] At present, these do not belong to the set of unary relations, but will shortly be added.

makes it possible to quickly retrieve the relevant Word2's and ascertain if these are involved in any other characteristic relations.

subject	num	sal	**object**	num	sal
price	316	22.8	victim	147	22.2
wicket	62	21.7	prey	51	18.2
rate	247	21.5	short	23	17.7
rain	155	21.4	foul	34	14.9
net	42	21.1	flat	29	12.5
profit	136	20.8	angel	15	11.2
snow	82	20.8	sick	18	9.2
dusk	39	20.6			
modifier	num	num	**particle**	num	sal
apart	335	335	over	638	16.9
short	247	247	off	738	16.8
ill	91	91	back	616	13.9
sharply	104	104	down	611	13.0
behind	78	78	by	98	12.5
headlong	22	22	through	127	12.4
dramatically	56	56	away	166	9.8
steadily	61	61	in	309	9.2
PP			**adj-comp**		
in love	867	44.0	asleep	604	26.2
into category	259	33.1	foul	98	30.0
into trap	142	31.3	silent	223	28.8
into disuse	69	28.0	short	142	26.6
into hand	143	26.8	due	79	25.4
by wayside	45	24.5	ill	109	22.3
on ear	47	23.9	vacant	34	18.7
out-of favour	36	23.2	open	44	12.4
to floor	106	23.2	**and-or**		
into step	39	23.0	rise	92	21.8
to knee	69	22.4	slip	22	14.2
into sleep	50	22.0	stumble	16	14.1
into place	88	21.8	trip	11	13.1
under spell	31	21.6	fall	34	12.9
into disrepair	26	21.6	stand	35	11.7
from grace	26	21.3	break	17	9.7
			hit	10	9.2

Table 3: Extract of Word Sketch for *fall* (v), BNC frequency = 23,836

The examples above involved combining a relation between Word1 and Word2, with characteristic unary relations on Word2. Another possibility would be cases where we could combine unary relations on Word1. Extending the principle further we could look for all significant patterns for Word1 or Word2, possibly introducing a new lexeme. Consider the reduced Word Sketch for the adjective *hot* in Table 4. The pattern **modifies bun** is at first rather mysterious.

Why should "hot bun" be such a strong collocational pattern? A glance at the examples reveals that it is of course that peculiar Easter delicacy the "hot cross bun" that creates this strong pattern. This can be automatically found by looking for characteristic patterns for the Word2 *bun* when occurring in this collocation, revealing that they are nearly all also modified by *cross*, allowing the collocation to be correctly identified and presented as **hot cross bun**.

subject	**num**	**sal**	**modifies**	**num**	**sal**	**modifies (cont.)**	**num**	**sal**
sun	34	26.1	water	976	31.0	drink	105	18.8
soup	8	11.2	bun	51	23.4	chocolate	60	18.8
weather	21	10.8	summer	196	23.1	sun	86	18.7
summer	10	10.2	cylinder	76	22.4	pursuit	61	18.1
iron	8	9.8	bath	97	21.1	tea	73	17.7
day	24	9.8	air	242	19.9	spot	102	16.8
water	18	8.8	balloon	52	19.3	spring	72	16.8
afternoon	6	7.5	weather	140	19.1	grill	24	16.3
it	552	7.3	flush	41	19.0	tap	37	16.0
PP			**adj-comp-of**			**and-or**		
on heel	42	24.0	serve	64	22.9	cold	257	23.9
under collar	21	20.5	pipe	14	17.4	humid	33	20.1
off press	12	16.7	blow	15	13.6	dry	114	19.5
on trail	9	14.0	scald	7	13.3	sweaty	24	16.7
with embarrassment	7	10.2	get	162	11.4	red	159	16.3
with rice	4	9.4	burn	11	10.6	sunny	37	15.8
with sauce	4	8.7	follow	8	7.4	boiling	22	15.8
against her	4	7.3	grow	29	7.2	sticky	29	15.6
for comfort	5	7.1	scorch	2	5.3	soapy	13	14.5

Table 4: Extract of Word Sketch for *hot* (adj), BNC frequency=9086

Similarly, if **hot cake** is a salient collocation, which it is although outside the range shown in the extract, then we should also be able to find "sell like hot cakes" by this method, merely by the fact that *cake* in this pattern, as well as being overwhelmingly plural, will also feature in the pattern **PP-inv sell like**.

This section has shown how combining patterns allows us to refine the collocations found without committing us to computationally expensive searches of all combinations of patterns in the corpus.

2.6. Future Developments

As noted above, we are envisaging modest extensions to the repertoire of grammatical relations, including unary relations, in order to increase the expressivity particularly when combining patterns.

We shall be adding automatically-induced thesaural categories [Lin1998] to the workbench, which will allow the compaction of patterns by generalising over the identity of Word2. As an illustration this will allow us to generalise the patterns **Bank of England**, **Bank of Scotland**, **Bank of Japan** etc. in the Word Sketch of *bank* to **bank of COUNTRY**.

We are also currently investigating the potential for using web data, with pages being downloaded and fed directly into the workbench. This strategy would extend the potential of the workbench beyond languages where large corpora are readily available.

3. Lexicographic evaluation

For the last two years, a set of 6000 Word Sketches has been used to compile the Macmillan Dictionary of English [Rundell2002], a new dictionary for advanced learners, with a team of thirty professional lexicographers using them every day, for every medium-to-high frequency noun, verb and adjective of English. The feedback we have received is that they are very useful, and change the way the lexicographer uses the corpus. They reduce the amount of time the lexicographers need to spend reading individual instances, and give the dictionary improved claims to completeness, as common patterns are far less likely to be missed. They provide lexicographers with plenty of examples to choose from, for editing and putting in the dictionary. This is all most popular with the project management.

4. Word Sketches and Word Sense Disambiguation

Word Sketches are designed to support lexicographic analysis. Where a word is polysemous, central to the analysis is the division of the word's semantic range into distinct meanings. An intimately related language-technology task is 'word sense disambiguation' (WSD): automatically working out which of a polysemous word's meanings applies, given a particular instance of use of the word. While WSD has made great progress in the last ten years, mostly through the application of machine-learning techniques and the use of large corpora, it now seems unlikely it can make much further progress unless it looks more closely at lexicography. Almost all large-scale WSD work to date has aimed to disambiguate between the meanings provided in an off-the-shelf resource, either a publisher's dictionary or WordNet.[9] It is increasingly apparent that the limits of this approach have been reached, as the analyses of polysemy in these resources are not sufficiently precise or explicit, to give the computer leverage to perform

[9] http://www.cogsci.princeton.edu/~wn

any better. What is needed is more explicit lexicography, coupled with good analyses.

In WASPS, the project of which Word Sketch development has been one part, we have developed a system in which a lexicographer not only analyses the word's meaning (starting from the Word Sketch), but also records the details of the analysis in a way which allows a WSD algorithm to accurately disambiguate new instances of the word. The system has performed well in WSD evaluation tasks. See [Kilgarriff and Tugwell2001] for fuller details.

5. Conclusion

A Word Sketch is an automatically-produced summary of a word's behaviour. They can be built from any large corpus, provided there are part-of-speech taggers, lemmatisers and grammars available for the language of the corpus. In this paper we have described how Word Sketches were built for English using the British National Corpus, how they were integrated into a lexicographer's workstation (which gave hyperlinked access to corpus examples). The Word Sketches have recently been used in a large scale dictionary project and have received favourable reviews.

Word Sketches build on recent developments in lexicography, corpus linguistics and Natural Language Processing to provide an improved way for lexicographers to find out what the corpus has to say about a word. In doing this, they ride a wave set in motion by Sue Atkins. We hope they contribute to better language description, and are thereby worthy followers in the Atkins tradition.

Bibliography

[Church and Hanks1989] Kenneth Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting*, Vancouver, Canada. Pages 76-83.

[Dunning1993] Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61-74.

[Evert and Krenn2001] Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In ACL *Proceedings of 39th Annual Meeting*, Toulouse, France. Pages 188-195.

[Hindle1990] Donald Hindle. 1990. Noun classification from predicate-argument structures. In *ACL Proceedings, 28th Annual Meeting*, Pittsburgh. Pages 268-275.

[Kilgarriff and Tugwell2001] Adam Kilgarriff and David Tugwell. 2001. Wasp-bench: an mt lexicographer's workstation supporting state-of-the-art lexical disambiguation. In *Proc. MT Summit VIII*, Santiago de Compostela, Spain, September. Pp 187-190.

[Kilgarriff1996] Adam Kilgarriff. 1996. Which words are particularly characteristic of a text? a survey of statistical approaches. In *Language Engineering for Document Analysis and Recognition*, pages 33-40, Brighton, England, April. AISB Workshop Series.

[Lin1998] Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL Proceedings*, pages 768-774, Montreal.

[Minnen et al.2000] Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proc. 1st Intnl. Conf. on Natural Language Generation*, pages 201-208, Mitzpe Ramon, Israel, June.

[Pedersen1996] Ted Pedersen. 1996. Fishing for exactness. In *Proc. Conf. South-Central SAS Users Group*, Texas.

[Rundell2002] Michael Rundell, editor. 2002. *Macmillan Dictionary of English for Advanced Learners*. Macmillan, London.

[Schulze and Christ1994] Bruno Schulze and Oliver Christ, 1994. *The IMS Corpus Workbench*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

[Tapanainen and Järvinen1998] Pasi Tapanainen and Timo Järvinen. 1998. Dependency concordances. *Int. Journal of Lexicography*, 11(3):187-204.