

Peeking into the Future of Writing

Jason S. Chang

Department of Computer Science
National Tsing Hua University, TAIWAN

2015-0306 Friday 11:00-12:30
Waseda University, Tokyo



Peeking into
the

Future of Writing

From Word Processor to

Jason S. Chang

Department of Computer Science
National Tsing Hua University, TAIWAN

2015-0306 Friday 11:00-12:30
Waseda University, Tokyo



Peeking into
the

Future of Writing

From Word Processor to

Interactive Writing Environment:

Jason S. Chang

Department of Computer Science
National Tsing Hua University, TAIWAN

2015-0306 Friday 11:00-12:30
Waseda University, Tokyo



Peeking into
the

Future of Writing

From Word Processor to

Interactive Writing Environment:

WriteAway

Jason S. Chang

Department of Computer Science
National Tsing Hua University, TAIWAN

2015-0306 Friday 11:00-12:30
Waseda University, Tokyo

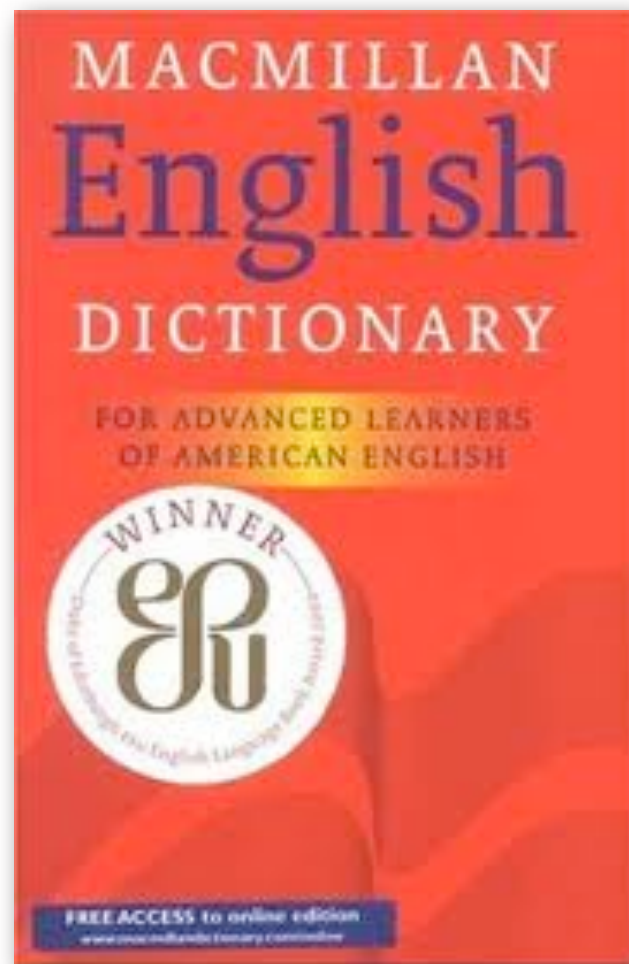


Dictionaries (and Grammar Books)

- For language learning, **dictionaries** have a long and honorable history
 - **Macmillan English Dictionary for Advanced Learners** (MEDAL)
 - Search the word *method*
- Recently, researchers (e.g., Sinclair) advocated
 - **Corpus** linguistics
 - **Corpus**-based lexicography,
 - Using a **concordance** in language teaching
 - E.g., **AntConc** with examples in *Citeseerx*
 - Small (.3%) sample (1.6 million words/0.7 million sentences)
 - From *Citeseerx* dataset (460 million words/20 million sentences)
 - Search the word *difficulty* -> almost as fast as **Google**
 - *But, what if we search the full Citeseerx?*

Macmillan English Dictionary

- Emphasis on Academic English
- Special Writing Section: Improve Your Writing Skills (IYWS)
 - Rhetoric Functions
 - Common Grammatical Errors



IMPROVE YOUR WRITING SKILLS	
Contents	
Introduction.....	W2
Writing Sections	
A. Adding Information.....	W4
B. Comparing and Contrasting: Describing similarities and differences.....	W5
C. Exemplification: Introducing examples.....	W9
D. Expressing Cause and Effect.....	W11
E. Expressing Personal Opinions.....	W15
F. Expressing Possibility and Certainty.....	W16
G. Introducing a Concession.....	W19
H. Introducing Topics and Related Ideas.....	W20
I. Listing Items.....	W23
J. Reformulation: Paraphrasing or clarifying.....	W24
K. Quoting and Reporting.....	W25
L. Summarizing and Drawing Conclusions.....	W28
Grammar Sections	
M. Articles.....	W29
N. Complementation: Patterns used with verbs, nouns and adjectives.....	W34
O. Countable and Uncountable Nouns.....	W38
P. Punctuation.....	W40
Q. Quantifiers.....	W43
R. Spelling.....	W46



Sylviane Granger

On a good day ...

method	Search
Dictionary	Thesaurus

MACMILLAN
DICTIONARY

- **a way of doing something**, especially a planned or established way
 - *It was a handmade rug produced by traditional methods.*
 - *Farming methods haven't changed here for decades.*
- **method of**: *We are trying to develop new methods of pollution control.*
- **method of doing something**: *They have adopted an alternative method of financing the scheme.*
- **method for doing something**: *Vaccination is the most effective method for preventing disease.*
- Collocations
 - **adj + method**: *effective, efficient, preferred, principal, reliable*
 - **v. + method**: *adopt, apply, choose, develop, devise, employ, pioneer, provide, use*

- Dictionaries have concise and organized information
 - **have difficulty with something:** *She's having difficulty with her schoolwork this year.*
 - **have difficulty (in) doing something:** *Six months after the accident, he still has difficulty walking.*
 - **great/considerable difficulty:** *We had considerable difficulty finding anywhere to park.*
 - **do something with/without difficulty:** *Seb was speaking with great difficulty.*
- But concordance has other advantages
 - Bring Your Own Data (BYOD)
 - Build Your Own Dictionary (BYOD)
 - Any texts, words, examples, in rich contexts
- Cobb (1998) reports that in vocabulary learning
 - Students used the concordance and made their own personal dictionaries (notes)
 - Learned and retained better

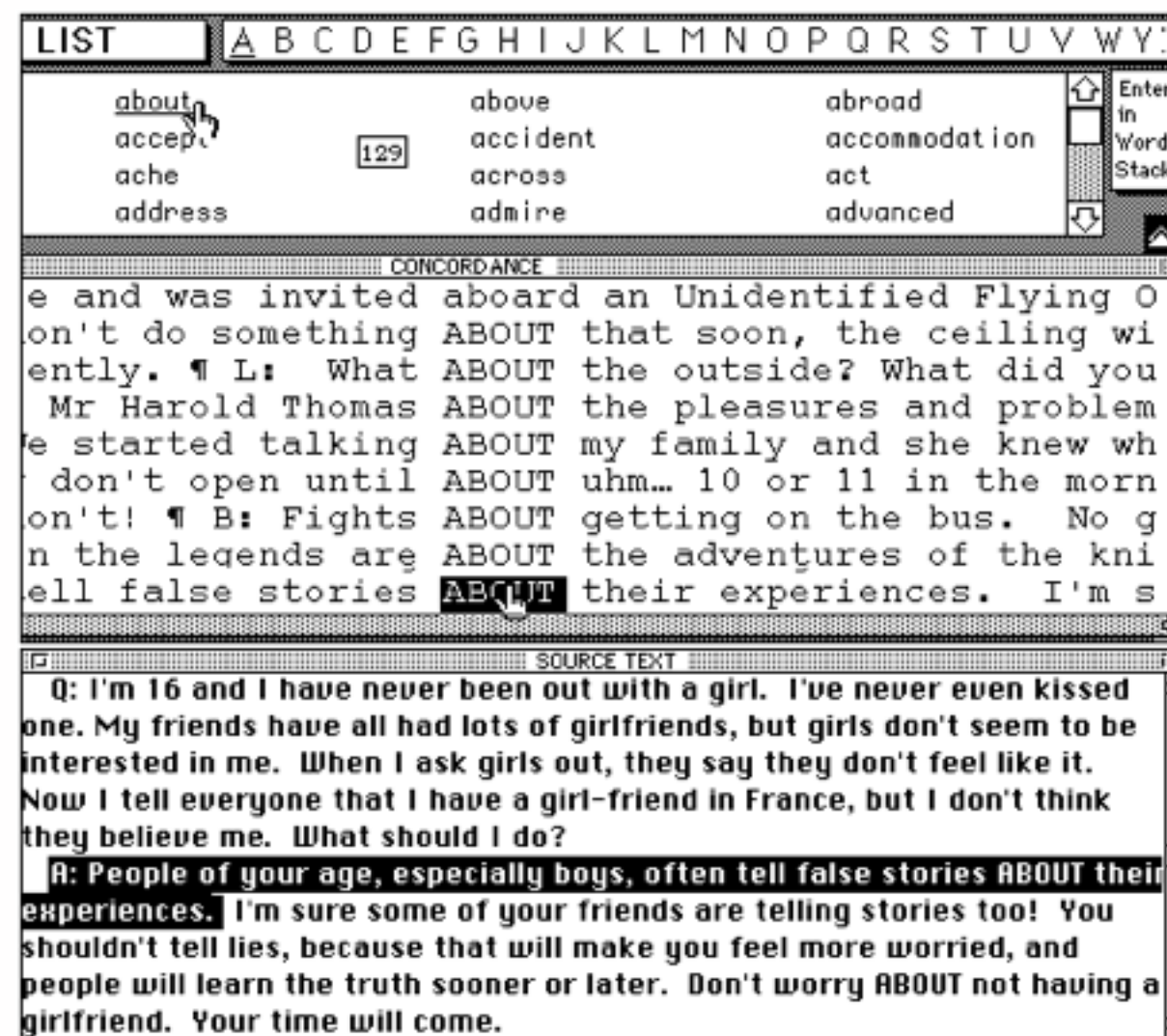
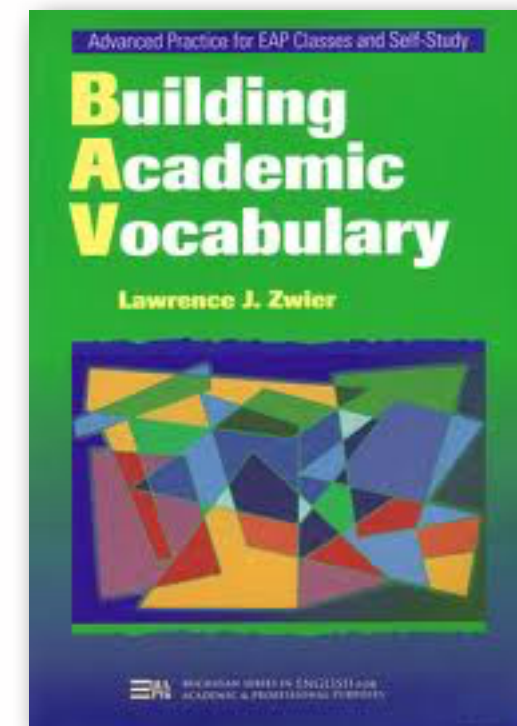
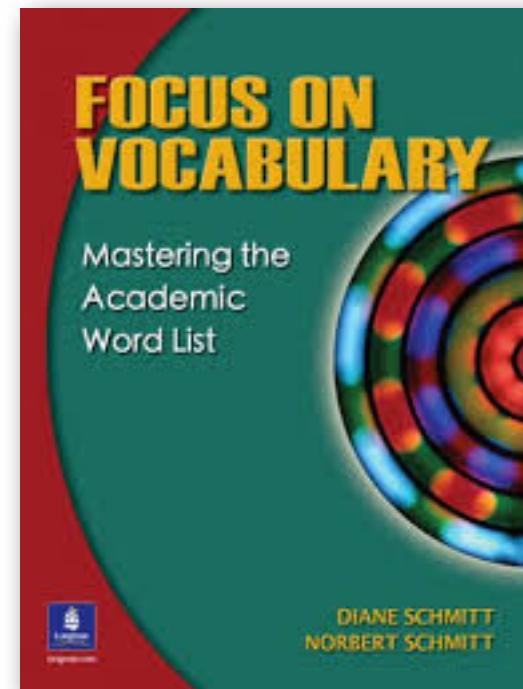
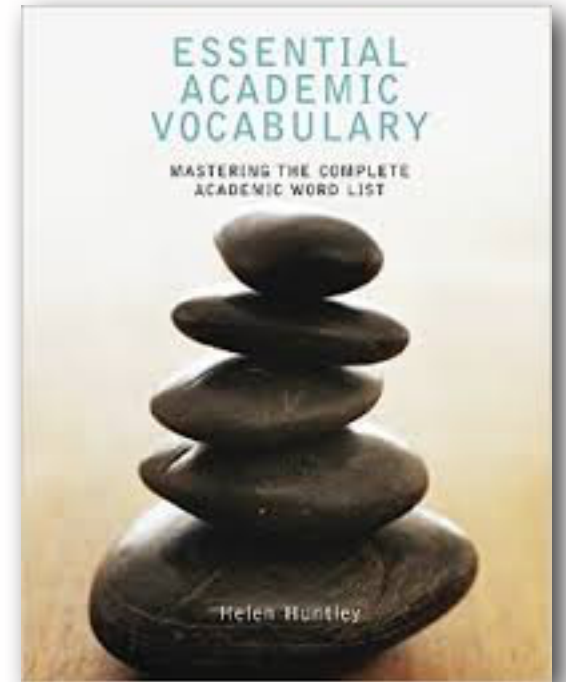


Figure 2. PET•2000 interface

Writing vs. Vocabulary Learning

- Why learning by *preparing* to do something?
 - Slow and indirect
- Why not learn by doing (or writing and speaking)
 - Learn vocabulary as you go
- In teaching Academic Writing, we should focus on
 - Academic Word List



Writing vs. Vocabulary Learning

- Why learning by *preparing* to do something?
 - Slow and indirect
- Why not learn by doing (or writing and speaking)
 - Learn vocabulary as you go
- In teaching Academic Writing, we should focus on
 - Academic Word List
 - OR **Rhetoric functions in Writing**

IMPROVE YOUR WRITING SKILLS	
Contents	
Introduction.....	IW2
Writing Sections	
A. Adding Information.....	IW4
B. Comparing and Contrasting: Describing similarities and differences.....	IW5
C. Exemplification: Introducing examples.....	IW9
D. Expressing Cause and Effect.....	IW11
E. Expressing Personal Opinions.....	IW15
F. Expressing Possibility and Certainty.....	IW16
G. Introducing a Concession.....	IW19
H. Introducing Topics and Related Ideas.....	IW20
I. Listing Items.....	IW23
J. Reformulation: Paraphrasing or clarifying.....	IW24
K. Quoting and Reporting.....	IW25
L. Summarizing and Drawing Conclusions.....	IW28
Grammar Sections	
M. Articles.....	IW29
N. Complementation: Patterns used with verbs, nouns and adjectives.....	IW34
O. Countable and Uncountable Nouns.....	IW38
P. Punctuation.....	IW40
Q. Quantifiers.....	IW43
R. Spelling.....	IW46

AntConc search for *difficulty* in CiteSeerX-gdex

AntConc 3.4.3m (Macintosh OS X) 2014

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 141

Hit	KWIC	File
1	of interest but due to mathematical difficulty . After discussing reasons why Java	gdex.sents.t
2	not well established due to technical difficulty and ambiguity of functional merit .	gdex.sents.t
3	classification of problems by their intrinsic difficulty and an understanding of what makes	gdex.sents.t
4	advanced nanotechnologies could (with great difficulty and little incentive) be used to	gdex.sents.t
5	e statistics students experience considerable difficulty applying their classroom knowledge	gdex.sents.t
6	spatial layout in hierarchical scales . A difficulty arises from the fact that the	gdex.sents.t
7	istent data representation , missing data and difficulty around understanding relationships	gdex.sents.t
8	stage amplifier , there is a relative difficulty because of the large number of	gdex.sents.t
9	endent . The proposed approach addresses this difficulty by reducing the verification problem	gdex.sents.t
10	formance . In such situations , metacognitive difficulty can improve brand evaluations becau	gdex.sents.t
11	claim . This will avoid the inevitable difficulty caused by an annual article-by	gdex.sents.t
12	in specific applications where operators have difficulty diagnosing ... Most object-oriented	gdex.sents.t
13	the system's segmentation algorithm has difficulty distinguishing between noise and sp	gdex.sents.t
14	of qualitative simulation techniques is the difficulty encountered when developing a quali	gdex.sents.t
15	analysis . The students initially experienced difficulty envisioning the investigation proce	gdex.sents.t
16	paradigm . 1 Introduction Perhaps the largest difficulty facing storage architects and perfor	gdex.sents.t
17	tory search sessions , users often experience difficulty finding and re-finding high-quality	gdex.sents.t

Search Term ☒ Words ☐ Case ☐ Regex

Search Window Size 50

difficulty Advanced

Start Stop Sort

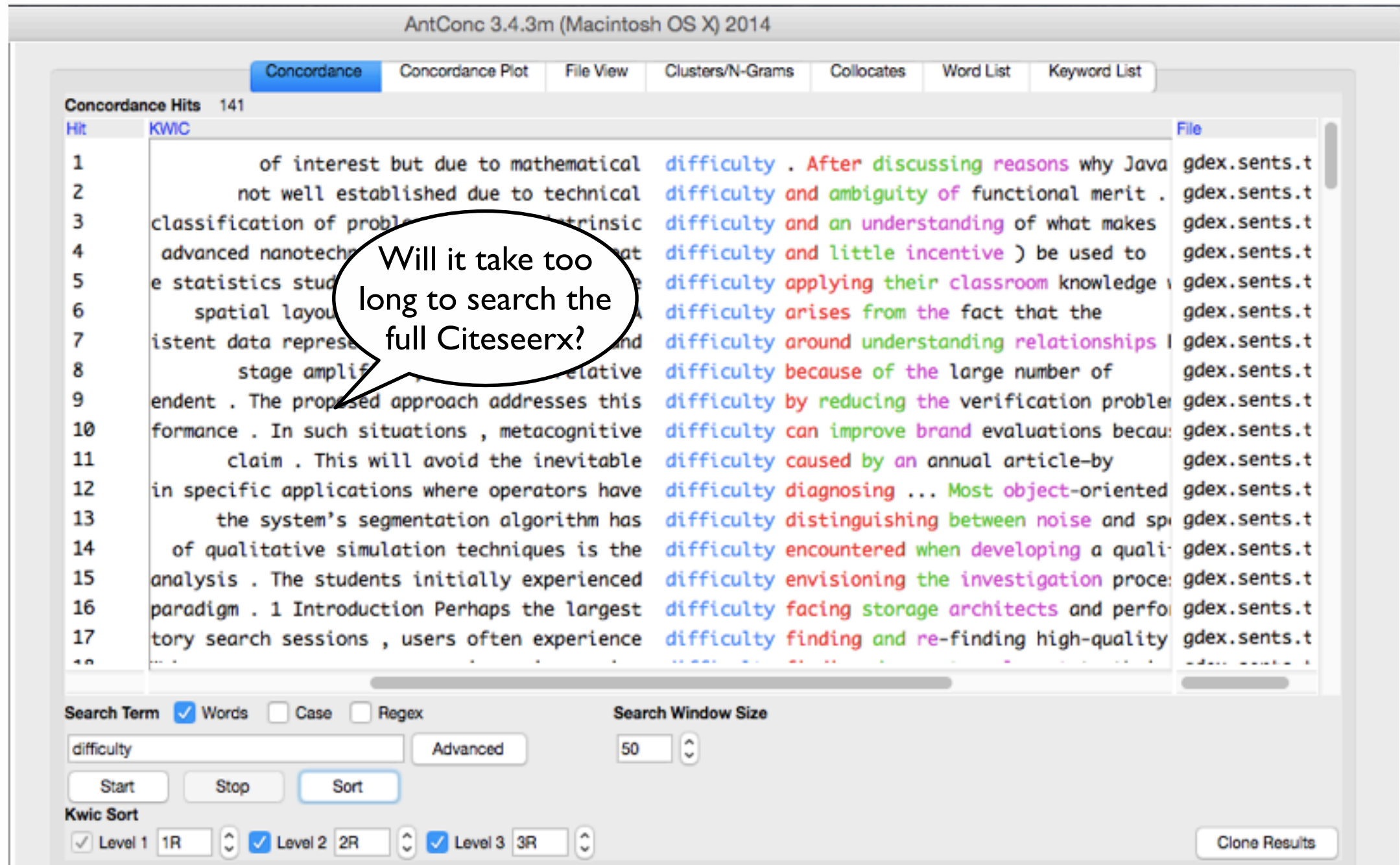
Kwic Sort

☒ Level 1 1R ☒ Level 2 2R ☒ Level 3 3R

Clone Results



AntConc search for *difficulty* in *CiteseerX-gdex*



AntConc search for *difficulty* in CiteSeerX-gdex

Is something missing in this small sample?

Will it take too long to search the full CiteSeerX?

Concordance Hits 141

Hit	KWIC	File
1	of interest but due to mathematical difficulty . After discussing reasons why Java	gdex.sents.t
2	not well established due to technical difficulty and ambiguity of functional merit .	gdex.sents.t
3	classification of problem . Intrinsic difficulty and an understanding of what makes	gdex.sents.t
4	advanced nanotechnology . That difficulty and little incentive) be used to	gdex.sents.t
5	e statistics study . The difficulty applying their classroom knowledge	gdex.sents.t
6	spatial layout . A difficulty arises from the fact that the	gdex.sents.t
7	istent data representation . And difficulty around understanding relationships	gdex.sents.t
8	stage amplification . The relative difficulty because of the large number of	gdex.sents.t
9	endent . The proposed approach addresses this difficulty by reducing the verification problem	gdex.sents.t
10	formance . In such situations , metacognitive difficulty can improve brand evaluations because	gdex.sents.t
11	claim . This will avoid the inevitable difficulty caused by an annual article-by	gdex.sents.t
12	in specific applications where operators have difficulty diagnosing ... Most object-oriented	gdex.sents.t
13	the system's segmentation algorithm has difficulty distinguishing between noise and sp	gdex.sents.t
14	of qualitative simulation techniques is the difficulty encountered when developing a quali	gdex.sents.t
15	analysis . The students initially experienced difficulty envisioning the investigation process	gdex.sents.t
16	paradigm . 1 Introduction Perhaps the largest difficulty facing storage architects and perfor	gdex.sents.t
17	tory search sessions , users often experience difficulty finding and re-finding high-quality	gdex.sents.t

Search Term ☒ Words ☐ Case ☐ Regex

Search Window Size 50

Start Stop Sort

Kwic Sort

☒ Level 1 1R ☒ Level 2 2R ☒ Level 3 3R

Clone Results

AntConc search for *difficulty* in CiteSeerX-gdex

Is something missing in this small sample?

Is this the best way to order examples?

Will it take too long to search the full CiteSeerX?

Concordance Hits 141

Hit	KWIC	File
1	of interest but due to mathematical difficulty . After	gdex.sents.t
2	not well established due to technical difficulty and ambiguity of functional merit .	gdex.sents.t
3	classification of problem intrinsic difficulty and an understanding of what makes	gdex.sents.t
4	advanced nanotechnology that difficulty and little incentive) be used to	gdex.sents.t
5	e statistics study difficulty applying their classroom knowledge	gdex.sents.t
6	spatial layout difficulty arises from the fact that the	gdex.sents.t
7	istent data representation and difficulty around understanding relationships	gdex.sents.t
8	stage amplification relative difficulty because of the large number of	gdex.sents.t
9	endent . The proposed approach addresses this difficulty by reducing the verification problem	gdex.sents.t
10	formance . In such situations , metacognitive difficulty can improve brand evaluations because	gdex.sents.t
11	claim . This will avoid the inevitable difficulty caused by an annual article-by	gdex.sents.t
12	in specific applications where operators have difficulty diagnosing ... Most object-oriented	gdex.sents.t
13	the system's segmentation algorithm has difficulty distinguishing between noise and sp	gdex.sents.t
14	of qualitative simulation techniques is the difficulty encountered when developing a quali	gdex.sents.t
15	analysis . The students initially experienced difficulty envisioning the investigation process	gdex.sents.t
16	paradigm . 1 Introduction Perhaps the largest difficulty facing storage architects and perfor	gdex.sents.t
17	tory search sessions , users often experience difficulty finding and re-finding high-quality	gdex.sents.t

Search Term ☒ Words ☐ Case ☐ Regex

difficulty Advanced Search Window Size 50

Start Stop Sort

Kwic Sort

☒ Level 1 1R ☒ Level 2 2R ☒ Level 3 3R

Clone Results

AntConc search for *difficulty* in CiteSeerX-gdex

Is something missing in this small sample?

Is this the best way to order examples?

Will it take too long to search the full CiteSeerX?

Do I want to see so many examples?

Concordance Hits 141

Hit KWIC File

1 of interest but due to mathematical difficulty . After ... why Java gdex.sents.t

2 not well established due to technical difficulty and ambiguity of functional merit . gdex.sents.t

3 classification of problem intrinsic difficulty and an understanding of what makes gdex.sents.t

4 advanced nanotechnology that difficulty and little incentive) be used to gdex.sents.t

5 e statistics study the difficulty applying their classroom knowledge gdex.sents.t

6 spatial layout difficulty arises from the fact that the gdex.sents.t

7 istent data representation and difficulty around understanding gdex.sents.t

8 stage amplification relative difficulty because of the gdex.sents.t

9 endent . The proposed approach addresses this difficulty by reducing the gdex.sents.t

10 formance . In such situations , metacognitive difficulty can improve the gdex.sents.t

11 claim . This will avoid the inevitable difficulty caused by an ann gdex.sents.t

12 in specific applications where operators have difficulty diagnosing ... most object-oriented gdex.sents.t

13 the system's segmentation algorithm has difficulty distinguishing between noise and sp gdex.sents.t

14 of qualitative simulation techniques is the difficulty encountered when developing a quali gdex.sents.t

15 analysis . The students initially experienced difficulty envisioning the investigation proces gdex.sents.t

16 paradigm . 1 Introduction Perhaps the largest difficulty facing storage architects and perfor gdex.sents.t

17 tory search sessions , users often experience difficulty finding and re-finding high-quality gdex.sents.t

Search Term ☒ Words ☐ Case ☐ Regex Advanced Search Window Size 50

Start Stop Sort

Kwic Sort ☒ Level 1 1R ☒ Level 2 2R ☒ Level 3 3R Clone Results

Concordances have limitations

- Speed
 - Returning results with a speed comparable to *Google*
- Coverage
 - Scaling up to a Web-scale dataset
- Presentation
 - **Ordering** the search results (putting important information first)
 - **Organizing** the search results (grouping similarly information)

What if we can have

- The best of both worlds
 - Dictionary
 - Concordance
- Solution
 - A concordance that looks/works like a dictionary

Previous Solutions (Flavor Manner Process)

- AntConc (vanilla)
 - bring your own data (BYOD), vanilla, examples ordered by neighboring
- GDEX or ForBetterEnglish (set menu)
 - Web examples organized by collocation (one example per collocation)
- Pattern Dictionary of English Verbs from CPA (handmade)
 - BNC examples organized by verb pattern

ForBetterEnglish

ForBetterEnglish

Search

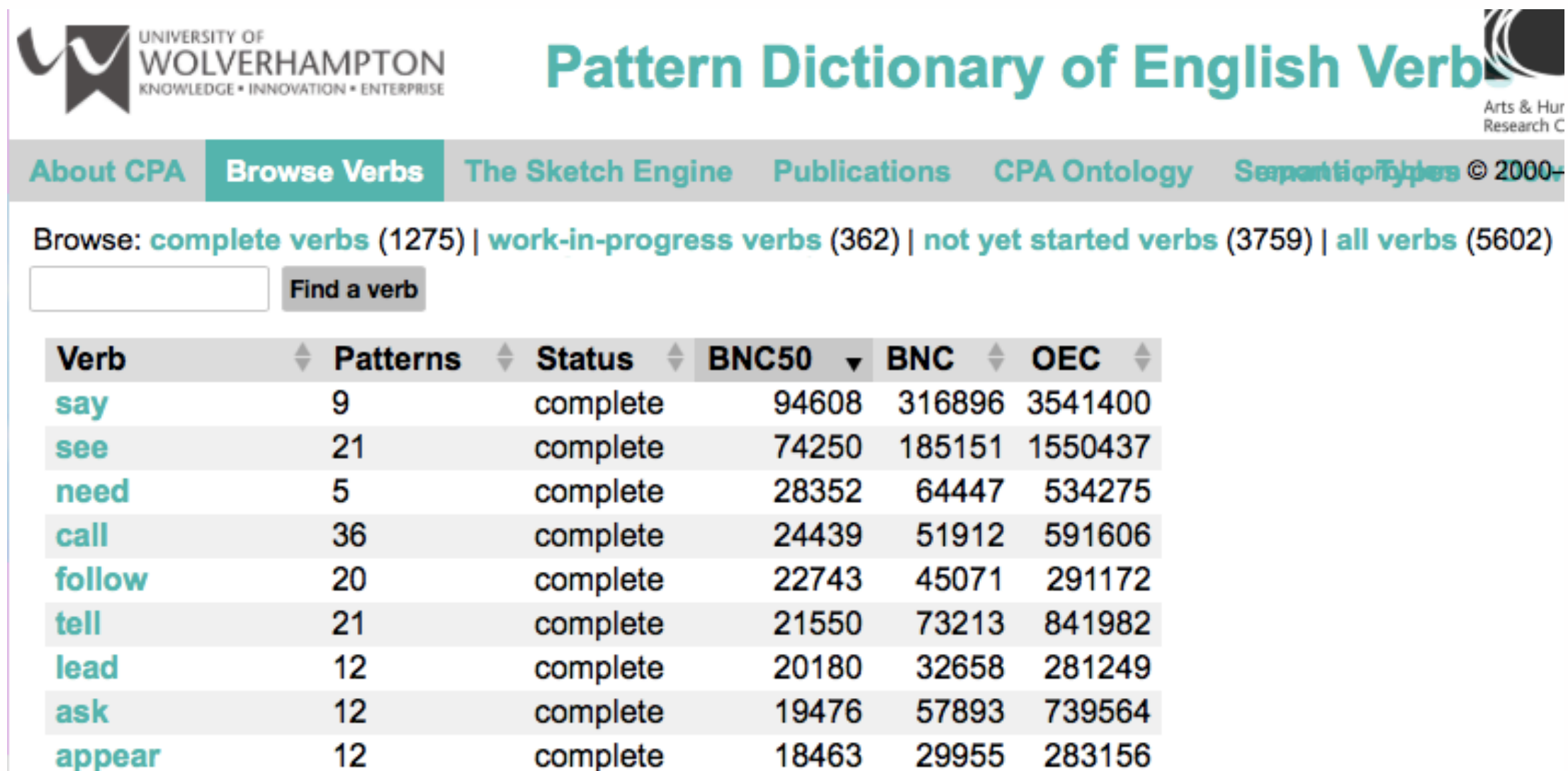
Sketch 文 Engine

abandon (v)

object	vehicle :	The official receiver should in no circumstances simply abandon a vehicle which is on the road .
	pretence :	And it was the day when Labour finally abandoned any pretence that it was serious about the reform of our public services .
	mine :	One of the attractions of caving , potholing and exploration of abandoned mines is its adventurous nature .
	ship :	With no way of repairing the damage , the order to abandon ship was given .
	quarry :	The dip of the rocks is clearly seen where slate is exposed in the now abandoned quarries and along the shore .
	siege :	In the Midlands , the Royalists abandon the siege of Warwick Castle on hearing news of the approach of the parliamentary relief force .
pp_on	land :	The vehicle is abandoned on private land - what do I do ?

Pattern Dictionary of English Verbs (CPA)

- Visit pdev.org.uk
- Choose from a list of 1275 (and growing) verbs
- View verb patterns with BNC concordance lines



The screenshot shows the homepage of the Pattern Dictionary of English Verbs (CPA) website. At the top left is the University of Wolverhampton logo with the tagline 'KNOWLEDGE • INNOVATION • ENTERPRISE'. The main title 'Pattern Dictionary of English Verbs' is in large teal letters. Below it is a navigation bar with links: 'About CPA', 'Browse Verbs' (highlighted), 'The Sketch Engine', 'Publications', 'CPA Ontology', and 'Semantic Types'. A copyright notice '© 2000-' is visible. Below the navigation bar, there is a search section with the text 'Browse: complete verbs (1275) | work-in-progress verbs (362) | not yet started verbs (3759) | all verbs (5602)'. A search box with the placeholder 'Find a verb' is present. Below this is a table of verbs with columns: Verb, Patterns, Status, BNC50, BNC, and OEC. The table lists verbs: say, see, need, call, follow, tell, lead, ask, and appear, each with its corresponding counts and status.

Verb	Patterns	Status	BNC50	BNC	OEC
say	9	complete	94608	316896	3541400
see	21	complete	74250	185151	1550437
need	5	complete	28352	64447	534275
call	36	complete	24439	51912	591606
follow	20	complete	22743	45071	291172
tell	21	complete	21550	73213	841982
lead	12	complete	20180	32658	281249
ask	12	complete	19476	57893	739564
appear	12	complete	18463	29955	283156

Example CPA Verb Pattern (abandon)

No.	%	Pattern / Implicature	
1	46%	[[Human Institution]] abandon [[Activity Plan]] [[Human Institution]] stops doing [[Activity]] or does not begin to do [[Plan]]	conc. exploit.
2	19%	[[Human Institution]] abandon [[Attitude]] [[Human Institution]] ceases to have [[Attitude]]	conc. exploit.
3	8%	[[Human {Human Group = Military}]] abandon [[Location]] [[Human {Human Group = Military}]] goes away from [[Location]] and does not live there any more, or (in military contexts) ceases to defend it	conc. exploit.
4	4%	[[Human]] abandon [[Artifact]] [[Human]] ceases to use or have possession of [[Artifact]] and leaves it somewhere at random	conc. exploit.
5	<1%	idiom [[Human]] abandon {NO DET ship} [[Human]] leaves {ship} at sea, by jumping into a life boat, life raft, or into the sea	conc. exploit.
6	14%	[[Human 1 Animal 1]] abandon [[Human 2 Animal 2]] (to [[Anything = Bad]]) [[Human 1 Animal 1]] goes away from and ceases to care for or look after [[Human 2 Animal 2]], with the result that [[Anything=Bad]] may get them or happen to them [[Human 1 Animal 1]] would normally be expected to look after and care for [[Human 2 Animal 2]] [[Human 2 Animal 2]] may be the offspring, wife (mate), or pet of [[Human 1 Animal 1]]	conc. exploit.
7	<1%	[[Human]] abandon [[Self]] {to [[Activity]] to [[Attitude]]} [[Human]] does [[Activity]] or adopts [[Attitude]] without thinking or caring about what is right, proper, or required by duty	conc. exploit.
8	<1%	[[Human]] abandon [[Self]] {to [[Deity]]} <i>Christianity.</i> [[Human]] gives up free will and allows [[Deity]] to make all decisions for him/her	conc. exploit.

© Patrick Hanks

CPA + BNC Example Sentences (Pattern 1)

1 pattern: **someone or somegroup abandon ACTIVITY/PLAN**
(e.g., **project, siege, development, idea, provision, plans**)

A07	. Catholic civil servants usually had to	abandon	1	any practical political project if they
A07	, James and his besiegers lost heart and	abandoned	1	the siege. </p><p> The mythical value of
A11	achieved and in 1983 the entire project was	abandoned	1	. </p><p> Had BR had more time and more funding
A18	the common people -- two ideas which were	abandoned	1	. And it should be noted that the actual
A2P	did not agree that Britain could or should	abandon	1	development, either for itself or for the
A31	After discreet soundings, they prudently	abandoned	1	the idea, which would have involved a major
A3G	Education Secretary, urging him not only to	abandon	1	the idea of additional CTCs but to hand
A46	Excise rates by 1993, the ministers agreed to	abandon	1	key provisions for revising VAT collection
A4J	Multatuli'). Moved to rectify this situation, he	abandoned	1	plans of working in the missionary field
A50	supporters when he confirmed the Government had	abandoned	1	plans for legislation to curb the activities
A5G	the Chancellor to alter but not altogether	abandon	1	the rule, effectively reducing the amount
A5R	the three who remained in France did not	abandon	1	their mission following the arrests in
A6B	primitive world for which, eventually, Celia	abandons	1	the Hollywood dream. Unreality and reality
A6G	countries, and never took effect; the USA itself	abandoned	1	the idea in 1946. Berle told Roosevelt
A77	Warden of the Cinque Ports. The Adjutant	abandons	1	the idea of a coffee and hurries towards
A7L	British filmmaking nosedived. UK filmmakers	abandoned	1	their innovations with film narrative,
A88	fees and other more radical ideas have been	abandoned	1	in the face of tough Treasury opposition
A8Y	scourge of the dozy British motorist, has not	abandoned	1	his life's work just because Her Indoors
A94	year Lloyd's of London insurance market is	abandoning	1	rules forcing underwriters to specialise

Concordances have limitations

- Limited Coverage
 - Even BNC (100 M) can be too small, not genre/domain specific
- Speed
 - Can we return results as fast as *Google* does?
- Presentation
 - **Ordering**: showing the most important usage first
 - **Organizing**: showing not all, but 2 or 3 examples per pattern

Concordances have limitations

- Limited Coverage
 - Even BNC (100 M) can be too small, not genre/domain specific
- Speed
 - Can we return results as fast as *Google* does?
- Presentation
 - **Ordering**: showing the most important usage first
 - **Organizing**: showing not all, but 2 or 3 examples per pattern
- Support linguistic search query
 - Show me the patterns after the word *difficulty*

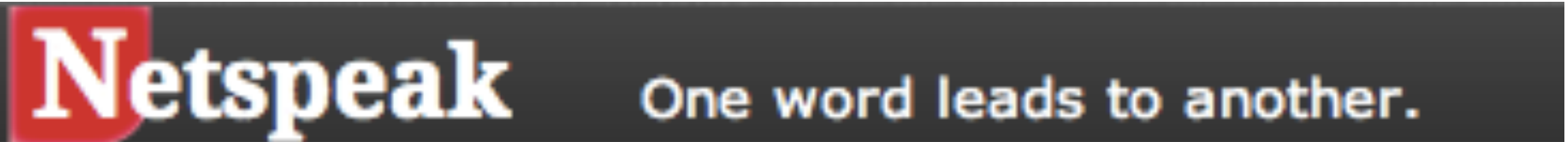
Concordances have limitations

- Limited Coverage
 - Even BNC (100 M) can be too small, not genre/domain specific
- Speed
 - Can we return results as fast as *Google* does?
- Presentation
 - **Ordering**: showing the most important usage first
 - **Organizing**: showing not all, but 2 or 3 examples per pattern
- Support linguistic search query
 - Show me the patterns after the word *difficulty*
- No query search
 - Show me the patterns after the word *difficulty*
 - *Automatically, proactively, unsolicited*

Solutions

- Web-scaled linguistic search engines: *NetSpeak* and *Linggle*
 - 1 trillion (10^{12}) words (noisy) Web page text
 - Index queries not keyword
 - Show search results as you type
 - Support linguistic queries
 - keywords, wildcard, *wild pos*, *synonyms*
 - Presentation: ordering examples by frequency
- Interactive Writing Environment: *WriteAway*
 - Search for information as you write
 - Group examples by *complementation pattern*
 - Order examples by *frequency*

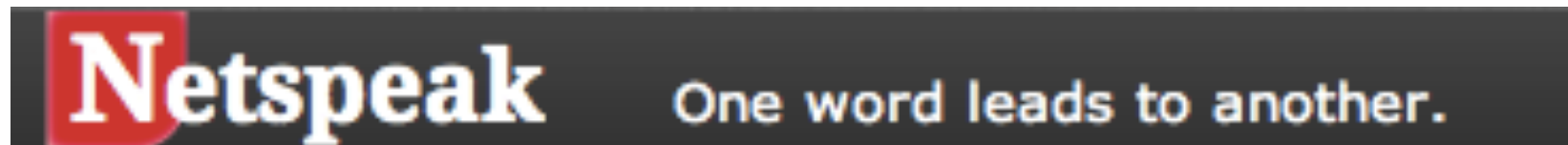
Netspeak.org – Find words in between and more ...



how to ? this
see ... works
it's [great well]
and knows #much
{ more show me }

The ? finds one word.
The ... find many words.
The [] compare options.
The # finds similar words.
The {} check the order. »

Netspeak.org – Find words in between and more ...



Find any number of words

Use dots, to find one, two, or more words at the same time.

waiting ... response			i x	Q
waiting for a response	42,000	62.0%	+	
waiting for response	9,600	14.1%	+	
waiting for your response	4,700	6.9%	+	
waiting for the response	4,300	6.3%	+	
waiting on a response	2,300	3.5%	+	
waiting for their response	1,200	1.7%	+	
waiting for his response	1,100	1.6%	+	
waiting for vendor response	940	1.4%	+	
waiting for my response	900	1.3%	+	
waiting for her response	710	1.0%	+	

demonstrating

demonstrating

Linggle

demonstrating
Linggle
Ling gle

demonstrating

Linggle

Linguistic Google

The One Minute Help

_	search for any word
~Term	search for the similar words of Term
~Term #N	the top N similar words of Term
?Term	search for Term optionally
*	match zero or more words
Term1 / Term2	with either Term1 or Term2
part-of-speech	search for word with specific part-of-speech. v. (verb), n. (noun), adj. (adjective), adv. (adverb) and prep. (preposition), det. (determiner), conj. (conjunction), pron. (pronoun), interj. (interjection)

demonstrating

demonstrating

WriteAway

demonstrating

WriteAway

Write Nonstop

demonstrating

Write Away

Write while being helped Nonstop

Writing programs TODAY

- People write programs in **Interactive Development Environment**
- In **IDE**
 - You type a few words (tokens)
 - The system will show you **suggestions**, what lies ahead
 - In other words, IDE help you with **prompts** and **autocomplete**
 - You **test drive** your program and get instant corrective feedback
 - **Syntax errors**
 - **Semantic errors**

My main point

- I will show you
 - **Future of Writing = Present of Programming**
 - **Future = Interactive Writing Environment**
- Why?
 - History repeats itself
 - We have better theory of the language
 - Lexical grammar
 - **Pattern Grammar**
 - We have much more **Data** to derive Pattern Grammar
 - **Statistical methods** are mature

Better Theory of Language

- **Old theory**: Two sides of a coin: **Vocabulary** + **Rules**
 - Words have parts of speech (POS)
 - Noun: object; Verb: action
 - Preposition/order: relation and event
 - Rules are related to POS, and almost independent of words
- **New theory**
 - Lexical grammar (e.g., Pattern Grammar)
 - Rules are intimately linked to words

Pattern Grammar

- PG is a model for describing the syntactic environments of individual words
- Each word has a set of patterns describing word usage in typical contexts
- One sense per pattern (often patterns are different for different word senses)

Sources:

- http://en.wikipedia.org/wiki/Pattern_grammar
- Hunston and Francis (2000): A corpus-driven approach to the lexical grammar of English

Pattern Grammar

Skim (v.) includes the following patterns in the COBUILD dictionary

- **V n off/from n:** *Skim the fat off the soup.* (limited prepositions allowed)
- **V n:** *Skim the wall surface smooth and get ready for painting.*
- **V over/across:** *Water skiers skimmed across the bay.*
- **V through n:** *Skim through the report and check for spelling mistakes?*

Sources:

- http://en.wikipedia.org/wiki/Pattern_grammar
- Hunston and Francis (2000): A corpus-driven approach to the lexical grammar of English

Dictionaries Embrace Pattern Grammar

MACMILLAN
DICTIONARY

Dictionaries Embrace Pattern Grammar

MACMILLAN DICTIONARY

- **have difficulty with something:** *She's having difficulty with her schoolwork this year.*
- **have difficulty (in) doing something:** *Six months after the accident, he still has difficulty walking.*
- **great/considerable difficulty:** *We had considerable difficulty finding anywhere to park.*
- **do something with/without difficulty:** *Seb was speaking with great difficulty.*

More Data Make It Easier to Derive PG

- **BNC** *British National Corpus* 100 million words
- **COCA** *Corpus of Contemporary American English* 450 million words
- **CiteseerX** *Scholarly Big Data* 460 million words (2.8G)
- **CTD** *Corpus of Taiwan Dissertations* 10 million words

Table 3: Collection and Usage Statistics

Statistic	Value
#Documents	3.5 million
#Unique documents	2.5 million
#Citations	80 million
#Authors	3-6 million
#docs added monthly	300,000
#docs downloaded monthly	300,000-2.5 million
Individual Users	800,000
Hits per day	2-4 million

Source: Wu, Zhaohui, et al. "Towards building a scholarly big data platform: Challenges, lessons and opportunities." in *Digital Libraries 2014*.

Increasing Effective NLP/Statistical Tools

- Parsing and shallow parsing (based on the old theory)
 - *Genia tagger* (Tsuruoka et al. 2005)
- Collocation Extraction
 - *XTract* (Smadja 1993)
- Extracting Good Dictionary Examples
 - *GDEX* (Kilgariff et al. 2008)
- Extracting Pattern Grammar from big data
 - *WriteAway* (Chang et al. 2015, in preparation)

Source: Tsuruoka, Yoshimasa, et al. "Developing a robust part-of-speech tagger for biomedical text." *Advances in informatics*. Springer Berlin Heidelberg, 2005. 382-392.

My main point

- I will show you
 - **Future of Writing = Present of Programming**
- Why?
 - History repeats itself
 - We have better theory of the language
 - Lexical grammar
 - Pattern Grammar
 - We have much more data to derive Pattern Grammar

NL and Statistical Processing of CiteSeerX (for WriteAway)

\$ Size

20,000,000 sentences; 460,000,000 words; 2.8 G bytes

\$ Sentence

Because of their deployment in critical applications , the dependability modeling and analysis of Multiple-Phased Systems is a task of primary relevance .

\$ Tagging (*Genia Tagger*)

— Part of speech tagging

- *IN IN PRP\$ NN IN JJ NNS , DT NN NN CC NN IN JJ NNS VBZ DT NN IN JJ NN .*
(prep, prep, possessive pronoun,)

— Base phrase tagging

- *I-PP H-PP I-NP H-NP H-PP I-NP H-NP O I-NP I-NP H-NP O H-NP H-PP I-NP H-NP H-VP I-NP H-NP H-PP I-NP H-NP O*
(PP-start, PP -start, NP-start, NP-end, PP, NP-start, NP-end, ...
PP = Prepositional Phrase, NP = Noun Phrase)

WriteAway is ...

WriteAway

We have difficulty|

less patterns

more patterns

less examples

more examples

[N] difficulty with something **755**

have difficulty with problems involving **25 5**

difficulty with this approach/method is **82 13**

[N] difficulty doing something **718**

have difficulty using it correctly **3 5**

Typewriter

WriteAway

Typewriter



We have difficulty|

less patterns

more patterns

less examples

more examples

[N] difficulty with something 755

have difficulty with problems involving 25 5

difficulty with this approach/method is 82 13

[N] difficulty doing something 718

have difficulty using it correctly 3 5

Typewriter + Dictionary

WriteAway

Typewriter



We have difficulty|

Dictionary



less patterns

more patterns

more examples

[N] difficulty with something **755**

have difficulty with problems involving **25 5**

difficulty with this approach/method is **82 13**

[N] difficulty doing something **718**

have difficulty using it correctly **3 5**

Typewriter + Dictionary + Concordance

WriteAway

Typewriter



We have difficulty|

less patterns

more patterns

more examples

Dictionary



[N] difficulty with something **755**

have difficulty with problems involving **25 5**

difficulty with this approach/method is

[N] difficulty doing something **718**

have difficulty using it correctly **3 5**

Concordance



Typewriter + Dictionary + Concordance = *WriteAway*

WriteAway

Typewriter



We have difficulty|

less patterns

more patterns

more examples

Dictionary



[N] difficulty with something **755**

have difficulty with problems involving **25 5**

difficulty with this approach/method is

[N] difficulty doing something **718**

have difficulty using it correctly **3 5**

Concordance



WriteAway

This paper

less patterns

more patterns

less examples

more examples

[N] paper does 123527

paper presents/proposes a 35626 11894

paper describes/discusses the 24462 7026

[N] paper does something 84532

paper describes/presents a system 1960 72

paper presents an approach to 1219 113

WriteAway

This paper presents|

less patterns

more patterns

less examples

more examples

[V] present something of something 52075

paper presents the results of a study conducted 630 7

paper presents the design and implementation of a mobile storage system called 259 6

[V] present something for doing something 27089

paper presents a parallel algorithm for solving the region growing problem based 235 7

we present a method for solving the following problem 142 11

This paper presents a method

[less patterns](#) [more patterns](#) [less examples](#) [more examples](#)

[N] method for something 25928

is a multistart type stochastic method for bound constrained global optimization problems 1330 6

by standard methods for one-dimensional systems 765 7

[N] method be 24538

method is applicable 17015 442

[N] method of something 19839

method of analysis/proof is 1325 115

on the method of moments 433 20

[N] method to do something 19659

of heuristic methods to solve hard computational search problems 528 8

presents a single-pass , view-dependent method to solve the general rendering equation 160 8

[N] method for doing something 18495

present a method for solving the following problem 960 11

of numerical methods for solving ordinary differential equations 326 5

WriteAway + Linggle ?

WriteAway

This paper discusses ?prep. * n.

less patterns

more patterns

less examples

more examples

English

英漢

[V] discuss something 36048

paper discusses an approach to
we discuss the issues involved
we discuss the problem of
we discuss the results and
we discuss what we have

WriteAway + *Linggle*

- Not yet
- But, entirely possible

Advantages of Using *WriteAway*

- Countability
 - Noticing the singular/plural forms in examples
- Verb tense/form
 - Noticing the verb forms in examples
- Article
 - Noticing the use or lack of articles
- Preposition
 - Noticing n.+prep. or prep.+n. patterns
 - E.g., evaluation on something
- Prep.+Verb Form
 - Noticing grammar patterns
 - E.g., method for doing something v.s. method to do something

What happens next?

- Improve *WriteAway*
- Enter a Competition for SVT Angels Emersion Grant to go to Silicon Valley
 - Talk Startup, Think Startup, and Breath Startup
- Get government funding
 - Taiwan Ministry of Economic Affairs (MOEA) as an Angel
- Start a KickStarter campaign
- Collaborate Partner Universities
 - Waseda University, National Taiwan University,
 - National Taiwan University of Science and Technology
- Come up with a business model: Lluber

What happens next?

- Improve **WriteAway**
 - Integrate into the environment of Google Doc, Microsoft Word, Gmail
 - Provide semantic pattern grammar
 - assist **someone** in something vs. assist something in something
 - attend **activity** vs. attend something
 - attend **institution** vs. attend something
 - Provide discourse related suggestion (7 rhetoric types, 200 pattern grammar rules)
 - **BKG** One of the **PROBLEM**: *One of the challenges/problems*
 - **GAP** However , it **NEED/CAUSE-PROB**: *However , it requires/remains*
 - **AIM** In this **WORK**, we **PRESENT**: *In this work , we present/propose/introduce*
 - **CTR** In **CONTRAST** , we **PRESENT**: *In contrast , we show*
 - **TXT** In this **TEXT_n** (,) we **PRESENT/ADDRESS**: *In this chapter (,) we present/describe/address*
 - Provide black-list pattern grammar
 - difficulty in doing something vs. * difficulty to do something
 - acquire knowledge vs. * learn knowledge
 - speak **LANGUAGE** adj vs. * speak **LANGUAGE** adv.

What happens next?

- Improve *WriteAway*
- Apply to SVTA Immersion Program to go to Silicon Valley
 - Talk Startup, Think Startup, and Breath Startup
- Get government funding
 - Taiwan Ministry of Economic Affairs (MOEA) as an Angel
- Start a KickStarter campaign
- Collaborate Partner Universities
 - Waseda University, National Taiwan University,
 - National Taiwan University of Science and Technology
- Come up with a business model: Lluber

U B E R + *TutorABC* = ?

- Brokerage is a major form of Web services
 - **Uber** matches up paying passengers with willing drivers where you can't get a taxi
 - uber**TAXI**, uber**X**, uber**XL**
 - Uber**BLACK**, Uber**SUV**
 - **TutorABC** matches up learners and tutors (conversation skills)
- How about Language Learning Uber (**Lluber**)
 - Match up learners and teachers (writing skills)
 - Robotic Tutors (*WriteAway*)
 - Fresh-and-blood Tutors (perhaps even foster Star Tutors and Fans)
- Diversify services
 - Tutors, editors, writing mates, teaching sessions, Writing As a Spectator Aport (WASS)
 - WriteAway-mediated collaborative writing
 - Online writers for hire

Embrace the Future of writing

with

Embrace the Future of writing with *Linggle*

Embrace
the Future of writing
with
Linggle
and
WriteAway