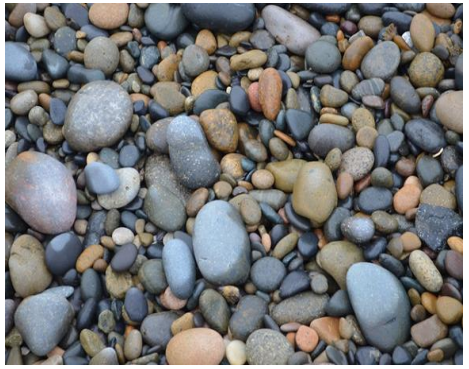
The quality of choices determines the quantity of Key words

PUNJAPORN P.
ANGVARRAH L.
22-6-2017

Why we need keyword analysis?

Keyword analysis

Target Corpus



VS

Benchmark Corpus

=

A keyword list

Key words vs keywords

keyword analysis
(two corpora + program)

↓

A keyword list
(keywords + keyness scores)

↓

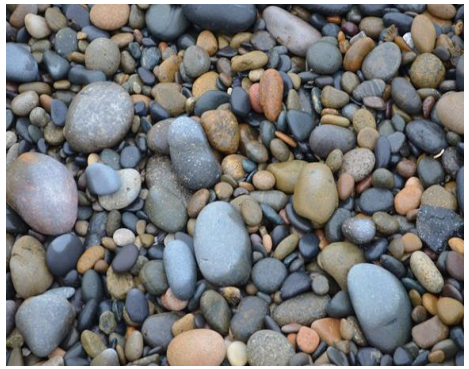
Key words
(Top words selected from a keyword list)

The problems

or

The problems

Target Corpus

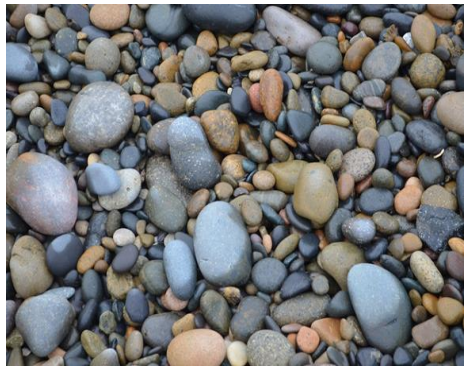


VS

Benchmark Corpus

=

Key words



VS

=

Factors influencing Key words

1 A target corpus (C1)

2 A benchmark corpus (C2)

3

rank	Word Type	Frequency in C1 (100,000 words)	Frequency in C2 (100,000,000 words)	Keyness scores
1	learners	698	689	7245.1
2	L2	359	63	4358.7
3	learning	660	9315	3833.2
4	study	763	22089	3393.7
5	language	632	19222	2752.3
6	students	566	14564	2643.4
7	proficiency	243	175	2623.1
8	test	446	14087	1910.6
9	research	494	26893	1616.7
10	learner	189	669	1579

Choices

1 A target corpus = **research article discussion**

VS

2 A benchmark corpus (C2)

=

3

Rank	Word Type	Frequency in C1 (100,000 words)	Frequency in C2 (100 m words)	Keyness scores
1	learners	698	689	7245.1
2	L2	359	63	4358.7
3	learning	660	9315	3833.2
4	study	763	22089	3393.7
5	language	632	19222	2752.3
6	students	566	14564	2643.4
7	proficiency	243	175	2623.1
8	test	446	14087	1910.6
9	research	494	26893	1616.7
10	learner	189	669	1579
11	teachers	346	11576	1444.8
12	english	439	23745	1441.9
13	findings	243	3278	1432.1
14	vocabulary	184	1226	1327.4
15	participants	202	2232	1266.6
16	knowledge	312	14548	1109.9
17	results	316	15348	1100.3
18	L1	107	145	1061.9
19	comprehension	134	605	1061.7
20	task	249	9162	995.4

20
60
100
400

Introduction (I)
Introduction, methodology, results (IMR)
A corpus of general English (BNC)

Top 25
Top 50
Top 100
Top 200

Aboutness

- Words indicating what a corpus is about

rank	KWs	keyness
1	be	406.6
2	may	294.6
3	that	210.2
4	**learners**	181.2
5	**findings**	169.6
6	more	158.8
7	this	139.0
8	**study**	138.5
9	should	133.3
10	future	132.5
11	**conclusion**	122.2
12	might	115.2
13	**finding**	109.4
14	can	108.0
15	**discussion**	92.3
16	**implications**	85.9
17	not	79.7
18	also	78.6
19	**limitations**	74.5
20	further	65.0

Numbers of texts in a target corpus

Study	Lists	No. of texts	Benchmark	Top n	Aboutness (%)	General Interpretation
1	1	20	(BNC)	(100)	73.0	No difference
	2	60	(BNC)	(100)	75.0	
	3	100	(BNC)	(100)	76.0	
	4	400	(BNC)	(100)	78.0	
			(IMR)	(200)		

Types of a benchmark corpus

Study	Lists	Benchmark	No. of texts	Top n	Aboutness (%)	General interpretation
2	1	I	(100)	(100)	43.0	Difference
	2	IMR	(100)	(100)	47.0	
	3	BNC	(100)	(100)	76.0	
			(60)	(50)		

Top words

Study	Lists	Top n	No. of texts	Benchmark	Aboutness (%)	General interpretation
3	1	25	(100)	(BNC)	96.0	Difference
	2	50	(100)	(BNC)	84.0	
	3	100	(100)	(BNC)	76.0	
	4	200	(100)	(BNC)	72.0	
			(60)			Difference
				(IMR)		No difference

Summary of findings

No	Factors in keyword analysis	Effects on #aboutness		Guidelines
1	Number of texts in D	✗		**20**, 60, 100, 400
2	Benchmark	✓	▸ #aboutness in D v BNC > D v I	**I, IMR, BNC**
3	Top words	✓ ✗	▸ BNC ▸ IMR	**Top 25**, 50, 100, 200