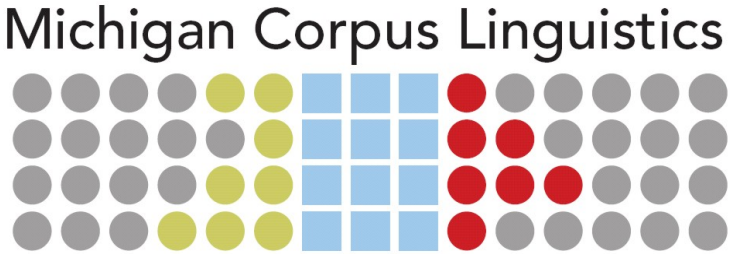


The Adjusted Frequency List:
Evaluating a method for producing cluster-sensitive frequency counts

Matthew Brook O'Donnell
University of Michigan

AACL Edmonton, Canada – 10 October 2009

How does a corpus linguist read a bedtime story?

the	34	a	5	papa	3
she	29	just	5	ran	3
in	14	said	5	second	3
and	13	bears	4	sitting	3
chair	10	down	4	tasted	3
porridge	10	into	4	then	3
bear	9	of	4	there	3
been	9	right	4	they	3
my	9	sleeping	4	ahhh	2
someone's	9	up	4	as	2
too	8	all	3	ate	2
was	8	baby	3	bedroom	2
goldilocks	7	bowl	3	big	2
it	7	but	3	came	2
this	7	eating	3	cried	2
to	7	exclaimed	3	decided	2
bed	6	first	3	forest	2
is	6	growled	3	from	2
so	6	lay	3	home	2
three	6	mama	3	last	2

How does a corpus linguist read a bedtime story?

and	13	bears	4	sitting	3
chair	10	down	4	tasted	3
porridge	10	into	4	then	3
bear	9	of	4	there	3
been	9	right	4	they	3
my	9	sleeping	4	ahhh	2
goldilocks	7	bowl	3	big	2
it	7	but	3	came	2
this	7	eating	3	cried	2
to	7	exclaimed	3	decided	2
bed	6	first	3	forest	2
is	6	growled	3	from	2
so	6	lay	3	home	2
three	6	mama	3	last	2

Answer #1.... type by type

But wait… we have clusters

clusters are dismantled in word list

in	14	said	5	second	3
and	13	bears	4	sitting	3
chair	10	down	4	tasted	3
porridge	10	into	4	then	3
bear	9	of	4	there	3
been	9	right	4	they	3
my	9	sleeping	4	ahhh	2
someone's	9	up	4	as	2
too	8	all	3	ate	2
was	8	baby	3	bedroom	2
goldilocks	7	bowl	3	big	2
it	7	but	3	came	2
this	7	eating	3	cried	2
to	7	exclaimed	3	decided	2
bed	6	first	3	forest	2
is	6	growled	3	from	2
so	6	lay	3	home	2
three	6	mama	3	last	2

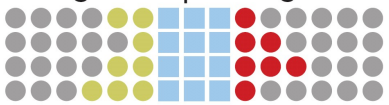
the	34	bed	6	all	3
she	29	in my	6	and she	3
in	14	is	6	baby	3
and	13	so	6	baby bear	3
chair	10	so she	6	been eating	3
porridge	10	three	6	been eating my	3
bear	9	a	5	been sitting	3
been	9	just	5	been sitting in	3
my	9	said	5	been sleeping	3
someone's	9	bears	4	been sleeping in	3
someone's been	9	down	4	bowl	3
too	8	into	4	but	3
was	8	is too	4	chair is	3
goldilocks	7	of	4	eating	3
in the	7	right	4	eating my	3
it	7	sleeping	4	eating my porridge	3
this	7	the three	4	exclaimed	3
to	7	the three bears	4	first	3
bear someone's	6	three bears	4	growled	3
bear someone's been	6	up	4	in my bed	3

How does a corpus linguist read a bedtime story?

Answer #2…. through words AND clusters

and	13	so	6	baby bear	3
chair	10	so she	6	been eating	3
porridge	10	three	6	been eating my	3
bear	9	a	5	been sitting	3
been	9	just	5	been sitting in	3
my	9	said	5	been sleeping	3
in the	7	right	4	eating my	3
it	7	sleeping	4	eating my porridge	3
this	7	the three	4	exclaimed	3
to	7	the three bears	4	first	3
bear someone's	6	three bears	4	growled	3
bear someone's been	6	up	4	in my bed	3

Michigan Corpus Linguistics

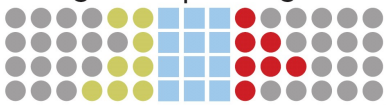


Motivation: Theoretical

- 'many chunks are as frequent as or more frequent than the single-word items which appear in the core vocabulary' (O'Keeffe, McCarthy and Carter 2006: 46)
- 'many words are frequent because of their strong constructional tendency' (Stubbs 2007)
- Phraseological Tendency – words 'go together and make meanings by their combinations' (Sinclair 2004: 29)

The frequency of individual words in corpora are often accounted for by the 'phraseological tendency', but this effect is masked in a simple frequency list where single-word items are inevitability elevated

Michigan Corpus Linguistics



Motivation: Practical

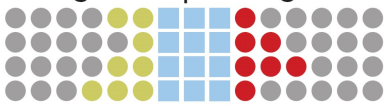
- Formulaicity project (http://ctr.elicorpora.info/formulaic-language-project)
  - Measuring frequency of n-grams across different corpora
    - Genre/register, text-type, text length, number of speakers, NS/NNS, learner levels
  - Used raw frequency thresholds
    - Over generate: on the other hand, on the, the other, other hand, on the other, the other hand etc.
  - Investigation of statistical/association measures
    - Complex to apply to n>2 (MI, Cost Critertion)
    - Iterative methods – Lexical Gravity Counts
  - Middle-way: Adjust simple frequency list to elevate rank of longer chunks.

Michigan Corpus Linguistics

Notions

- repeated chunks/clusters/n-gram are formulaic for language users, that is they represent a single choice
  - A **formulaic sequence** is "a sequence, continuous or discontinuous, of words, or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002: 9)
- Sinclair
  - Idiom principle (1987, 1991 etc.)
  - Linear Unit Grammar (2006)

Michigan Corpus Linguistics



Adjusted frequency list

- Designed for identification and quantification of chunks/formulas (e.g. *you know* in spoken corpora) using a (frequency) threshold
- If these are single choice items for speakers should be counted as single items and internal constituents left uncounted
  - BNC Baby Demographic:

	Freq	**Rank**
you	29688	2
know	7659	20
you know	3606	62

Michigan Corpus Linguistics

An Example

- Consider the following ‘text’, consisting of 5 types and 14 tokens:

A B C A B C A B D A E A B C

Michigan Corpus Linguistics

An Example

- The table below contains the frequency lists for all the 1, 2 and 3 grams in this text. The lists are ordered by frequency and then alphabetically.

A B C A B C A B D A E A B C

Words (1-grams)		2-grams		3-grams	
A	5	A B	4	A B C	3
B	4	B C	3	B C A	2
C	3	C A	2	C A B	2
D	1	A E	1	A B D	1
E	1	B D	1	A E A	1
		D A	1	B D A	1
		E A	1	D A E	1
				E A B	1

Michigan Corpus Linguistics

An Example

A B C A B C A B D A E A B C

- This produces a combined 1-3 gram list of 20 items:

Item	Freq	Item	Freq
A	5	A E	1
A B	4	A E A	1
B	4	B D	1
A B C	3	B D A	1
B C	3	D	1
C	3	D A	1
B C A	2	D A E	1
C A	2	E	1
C A B	2	E A	1
A B D	1	E A B	1

as N grows this list will grow and contain more and more overlap of parts of n-grams

N-gram list	**Types**
1	5
1 to 2	12
1 to 3	20
1 to 4	29
1 to 5	38
1 to 6	47
1 to 7	55
1 to 8	62
1 to 9	68

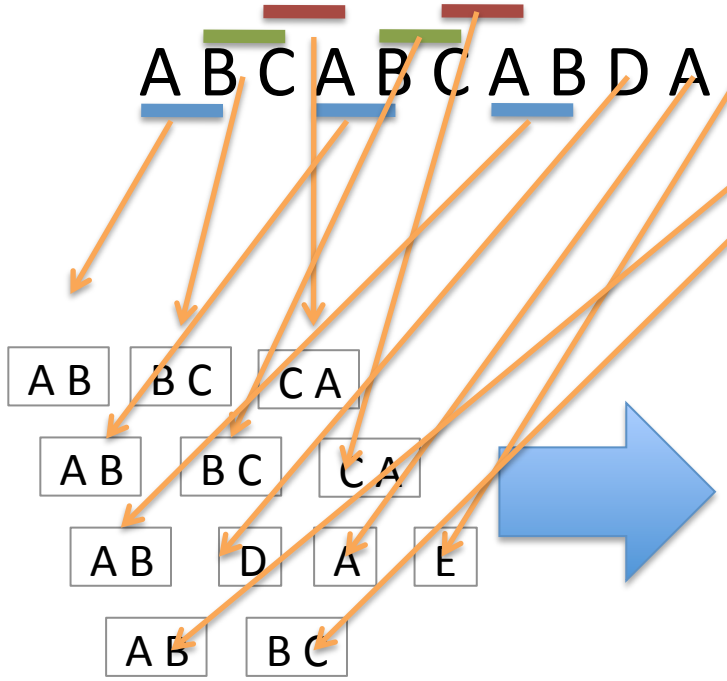
Michigan Corpus Linguistics

An Example

A B C A B C A B D A E A B C

- Consider the combined 1-2 gram list with 12 items with a frequency threshold of 2 for inclusion of 2-grams
- Create frequency list from text counting chunks as a single item and ignoring their component words

Rank	Item	Freq
1	A	5
2	A B	4
3	B	4
4	BC	3
5	C	3
6	C A	2
7	D	1
8	E	1



Rank	Item	Freq
1	A B	4
2	B C	3
3	C A	2
4	A	1
5	D	1
6	E	1

Michigan Corpus Linguistics

An Example

A B C A B C A B D A E A B C

Unadjusted

Rank	Item	Freq
1	A	5
2	A B	4
~~3~~	~~B~~	~~4~~
4	BC	3
~~5~~	~~C~~	~~3~~
6	C A	2
7	D	1
8	E	1

Adjusted

Rank	Item	Freq
1	A B	4
2	B C	3
3	C A	2
4	A	1
5	D	1
6	E	1

Consider:

A = *of*

B = *course*

A B = of course

Michigan Corpus Linguistics

An Example

A B C A B C A B D A E A B C

- Consider the combined 1-3 gram list with 11 items with a frequency threshold of 2 for inclusion of 2 & 3-grams
- Create frequency list from text counting chunks as a single item and ignoring their component words

Rank	Item	Freq
1	A	5
2	A B	4
3	B	4
4	ABC	3
5	BC	3
6	C	3
7	B C A	2
8	C A	2
9	C A B	2
10	D	1
11	E	1

A B C B C A C A B

A B C B C A C A B

A B D A E

A B C

Rank	Item	Freq
1	A B C	3
2	B C A	2
3	C A B	2
4	A	1
5	A B	1
6	D	1
7	E	1

Michigan Corpus Linguistics

An Example

A B C A B C A B D A E A B C

Unadjusted

Rank	Item	Freq
1	A	5
2	A B	4
~~3~~	~~B~~	~~4~~
4	ABC	3
~~5~~	~~BC~~	~~3~~
~~6~~	~~C~~	~~3~~
7	B C A	2
~~8~~	~~C A~~	~~2~~
9	C A B	2
10	D	1
11	E	1

Adjusted

Rank	Item	Freq
1	A B C	3
2	B C A	2
3	C A B	2
4	A	1
5	A B	1
6	D	1
7	E	1

Consider:

A = *I*
B = *don't*
C = know

A B = *I don't*
C A = *don't know*
A B C = I don't know

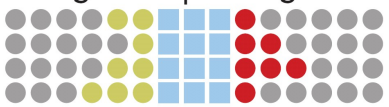
‘Three Bears’ Top 60 1-3 grams Unadjusted

the	34	bed	6	all	3
she	29	in my	6	and she	3
in	14	is	6	baby	3
and	13	so	6	baby bear	3
chair	10	so she	6	been eating	3
porridge	10	three	6	been eating my	3
bear	9	a	5	been sitting	3
been	9	just	5	been sitting in	3
my	9	said	5	been sleeping	3
someone's	9	bears	4	been sleeping in	3
someone's been	9	down	4	bowl	3
too	8	into	4	but	3
was	8	is too	4	chair is	3
goldilocks	7	of	4	eating	3
in the	7	right	4	eating my	3
it	7	sleeping	4	eating my porridge	3
this	7	the three	4	exclaimed	3
to	7	the three bears	4	first	3
bear someone's	6	three bears	4	growled	3
bear someone's been	6	up	4	in my bed	3

‘Three Bears’ Top 60 1-3 grams Adjusted (thres. 3+)

she	16	bed	3	she tasted the	3
and	10	been eating my	3	sitting in my	3
the	9	been sitting in	3	sleeping in my	3
goldilocks	7	been sleeping in	3	someone's been eating	3
in the	7	bowl	3	someone's been sitting	3
bear someone's been	6	but	3	someone's been sleeping	3
so she	6	eating my porridge	3	the first	3
a	5	exclaimed	3	the mama bear	3
was	5	growled	3	the second	3
chair	4	in my bed	3	then	3
down	4	in my chair	3	there	3
is too	4	into the	3	they	3
of	4	it all	3	this chair is	3
porridge	4	it was	3	this porridge is	3
the three bears	4	just right	3	to the	3
to	4	mama bear someone's	3	ahhh	2
too	4	papa bear	3	as	2
up	4	ran	3	ate	2
and she	3	said the mama	3	bedroom	2
baby bear	3	she lay	3	big	2

Michigan Corpus Linguistics



An interesting side-effect: Classification of instances?

Unadjusted: *just* occurs 5 times, *just right* occurs 3 times

```
1   "Ahhh, this porridge is just right," she said happily and
2   ir."Ahhh, this chair is just right," she sighed. But just
3   right," she sighed. But just as she settled down into the
4   he third bed and it was just right. Goldilocks fell aslee
5   !" exclaimed Baby bear. Just then, Goldilocks woke up an
```

Adjusted: *just right* occurs 3 times with attributive function

```
1   "Ahhh, this porridge is just right," she said happily and
2   ir."Ahhh, this chair is just right," she sighed. But just
3   he third bed and it was just right. Goldilocks fell aslee
```

Adjusted: *just* occurs 2 times with temporal function

```
1   right," she sighed. But just as she settled down into the
2   !" exclaimed Baby bear. Just then, Goldilocks woke up an
```

METHOD: Simple non-indexed version

1. Construct frequency lists (or a single combined list) for all items length 1 to *n* using the standard moving word window method and no frequency threshold (i.e. all items down to single occurrence).
2. Remove all items of length 2 to *n* that occur less than desired threshold used for formula/unit status.
3. For each remaining n-gram with frequency *f* (in descending order by length, i.e. *n* to 2) derive each of its component sub-items.
   - So for the trigram A B C there are bigrams A B and B C and three single items A B and C.
4. Reduce the frequency of each of these sub-items by *f*.

METHOD: Indexed version

1. Construct indexed frequency lists for all items length 1 to *n*, so that each instance of an item is recorded with reference its source file and position within that file (either just start or both start and end offsets).
2. Remove all items of length 2 to *n* that occur less than desired threshold used for formula/unit status.
3. For each remaining n-gram with frequency *f* (in descending order by length, i.e. *n* to 2) derive each of its component sub-items.
   - So for the trigram A B C there are bigrams A B and B C and three single items A B and C.
4. For each of the sub-items identified in Step 3, scan their index records for an occurrence that falls within the offset range the larger n-gram and remove record.

BNC Baby Demographic Top 40 1-3 grams Unadjusted

1	i	30371
2	you	29688
3	the	27698
4	it	21834
5	and	19845
6	a	19600
7	to	17180
8	that	14722
9	yeah	14303
10	oh	10398
11	in	10133
12	no	9804
13	of	9799
14	it's	8534
15	well	8478
16	what	8171
17	on	7951
18	is	7816
19	have	7802
20	know	7659
21	one	7488
22	do	7280
23	was	7133
24	got	6842
25	we	6686
26	he	6618
27	don't	6477
28	they	6475
29	but	6178
30	so	6148
31	there	6125
32	that's	5957
33	for	5673
34	mm	5662
35	not	5270
36	go	4941
37	be	4869
38	this	4781
39	get	4772
40	like	4744

BNC Baby Demographic Top 40 1-3 grams Adjusted

1	**i don't know**	1205
2	and	1088
3	the	647
4	**do you want**	544
5	**one two three**	521
6	**i don't think**	517
7	of	497
8	in	466
9	**two three four**	451
10	a	446
11	or	444
12	**in the**	417
13	**on the**	406
14	yeah	404
15	**a lot of**	395
16	er	381
17	**and the**	379
18	for	379
19	**for the**	369
20	**what do you**	365
21	your	365
22	erm	356
23	to	343
24	**no no no**	342
25	no	333
26	with	326
27	is	319
28	**you have to**	308
29	my	306
30	that	303
31	**of the**	293
32	**i mean i**	286
33	**you want to**	281
34	mm	279
35	oh	279
36	**with the**	277
37	it's	264
38	**a bit of**	262
39	his	262
40	on	260

Michigan Corpus Linguistics

BNC Baby Demographic

- 84 of the top 150 items in the adjusted list are bi- or tri-gram items, compared to 18 of the top 150 in the unadjusted list.
- The most frequent item in the adjusted list is the tri-gram *I don't know*.
- Many of the bi- and tri-gram chunks are central clause fragments for
  - questions (*do you want*, *what do you*, *do you think*, *have you got*, *do you know*, *can I have*)
  - directives (*have a look*, *you have to*, *you've got to*
  - declarative statements (*I don't know*, *I don't think*, *I think it's*, *I want to*).

MICASE

1	the	28936
2	you	17419
3	i	16816
4	and	15380
5	that	15352
6	to	13938
7	a	13522
8	of	13317
9	it	11111
10	is	10211
11	in	9412
12	so	7761
13	um	7604
14	this	6946
15	like	6852
16	yeah	6622
17	uh	6038
18	okay	5816
19	it's	5746
20	what	5583
21	have	5572
22	we	5507
23	know	4953
24	but	4714
25	be	4237
26	just	4212
27	right	4202
28	if	4097
29	do	4085
30	for	4021

1	yeah	3384
2	okay	3184
3	the	2772
4	mhm	2635
5	right	2280
6	you	2269
7	of	1968
8	and	1955
9	is	1438
10	that	1434
11	i	1410
12	um	1375
13	it	1342
14	a	1289
15	to	1283
16	in	1189
17	uh	1046
18	so	1005
19	or	949
20	like	885
21	know	850
22	this	845
23	have	824
24	it's	819
25	no	818
26	don't	804
27	was	765
28	think	732
29	what	728
30	xx	706

MICASE

31	on	3973
32	they	3962
33	was	3898
34	one	3863
35	are	3723
36	that's	3658
37	not	3501
38	or	3332
39	about	3294
40	think	3271
41	don't	3220
42	mhm	2922
43	with	2912
44	can	2900
45	there	2830
46	you know	2767
47	xx	2755
48	at	2731
49	well	2656
50	of the	2655
51	no	2603
52	oh	2488
53	in the	2366
54	all	2355
55	then	2345
56	as	2289
57	i'm	2244
58	mean	2180
59	would	2113
60	how	2075

31	oh	703
32	we	670
33	are	653
34	then	630
35	mean	608
36	for	606
37	just	596
38	on	581
39	not	572
40	of the	545
41	that's	543
42	they	540
43	do	532
44	yes	514
45	and the	509
46	can	504
47	with	479
48	one	468
49	uhuh	444
50	in the	438
51	there	437
52	you know	434
53	about	417
54	if	407
55	kind	403
56	but	399
57	would	399
58	be	397
59	alright	391
60	here	367

CHILDES – Naomi (14-23 months) Unadjusted

1	it	403
2	this	375
3	what's	371
4	i	321
5	what's this	295
6	down	172
7	want	166
8	daddy	137
9	mommy	136
10	get	124
11	that	118
12	more	115
13	no	110
14	the	104
15	i want	96
16	up	93
17	where's	91
18	is	88
19	baby	81
20	go	80
21	juice	75
22	on	75
23	look	74
24	going	71
25	oof	69
26	a	65
27	at	64
28	don't	64
29	look at	61
30	like	60
31	that's	59
32	there	58
33	book	57
34	what's that	56
35	kitty	55
36	you	55
37	birdie	53
38	like it	53
39	georgie	52
40	doggie	50
41	coffee	49
42	fall	48
43	where	47
44	do	46
45	fall down	46
46	sleeping	46
47	off	45
48	peanutbutter	43
49	recorder	43
50	take	43
51	i'm	42
52	can	41
53	don't like	41
54	in	39
55	moon	39
56	see	39
57	don't like it	38
58	more juice	38
59	oh	38
60	hi	36
61	sun	36
62	hot	35
63	nomi	35
64	my	34
65	oof oof	34
66	open	34
67	broken	33
68	i get	33
69	toast	33
70	at the	32
71	get up	32
72	red	32
73	take it	32
74	eight	31
75	read	31
76	shadow	31
77	ah	30
78	look at the	30
79	did	29
80	flower	29
81	sit	29
82	to	29
83	here	28
84	piggie	28
85	get down	27
86	get it	26
87	nightnight	26
88	nine	26
89	push	26
90	apple	25

CHILDES – Naomi (14-23 months) – Adjusted (5+)

1	what's this	295
2	it	133
3	mommy	110
4	no	110
5	daddy	100
6	baby	81
7	want	80
8	i	79
9	where's	58
10	is	57
11	what's that	56
12	on	55
13	going	54
14	down	51
15	get	51
16	go	51
17	like	51
18	doggie	50
19	up	50
20	a	49
21	kitty	49
22	more	49
23	the	46
24	georgie	45
25	this	43
26	sleeping	41
27	at	40
28	see	39
29	there	39
30	where	39
31	don't like it	38
32	oh	38
33	birdie	37
34	recorder	37
35	that	36
36	nomi	35
37	you	35
38	fall down	34
39	juice	34
40	oof oof	34
41	broken	33
42	i'm	33
43	more juice	33
44	book	32
45	coffee	31
46	hi	31
47	i want	31
48	peanutbutter	31
49	shadow	31
50	did	29
51	flower	29
52	in	29
53	here	28
54	piggie	28
55	moon	27
56	toast	27
57	open	26
58	red	26
59	that's	26
60	ah	25
61	apple	25
62	ball	25
63	diaper	25
64	get up	25
65	look at the	25
66	sun	25
67	my	23
68	eating	22
69	gammy	22
70	monkey	22
71	off	22
72	how	21
73	meow	21
74	need	21
75	duck	20
76	house	20
77	i did it	20
78	man	20
79	nightnight	20
80	another	19
81	cheerios	19
82	green	19
83	hot	19
84	i want it	19
85	tired	19
86	again	18
87	bug	18
88	banana	17
89	boo	17
90	find	17

CHILDES – Naomi (24-35 months) Unadjusted

1	i	996
2	it	516
3	that	470
4	the	453
5	a	397
6	you	376
7	this	340
8	and	320
9	uh	311
10	what's	293
11	my	283
12	on	248
13	want	237
14	is	236
15	in	229
16	there	217
17	yeah	203
18	one	194
19	to	190
20	me	179
21	no	179
22	down	168
23	up	168
24	i want	166
25	i'm	165
26	get	157
27	don't	154
28	go	154
29	here	154
30	what's that	143

31	what	138
32	do	132
33	put	125
34	going	123
35	that's	116
36	georgie	114
37	daddy	110
38	doing	109
39	where's	104
40	see	95
41	have	94
42	this is	94
43	like	92
44	nomi	90
45	mommy	86
46	it's	83
47	i don't	81
48	wanna	81
49	baby	78
50	oh	77
51	at	76
52	look	73
53	need	71
54	off	71
55	yup	71
56	two	70
57	are	68
58	another	67
59	can	66
60	draw	66

61	i need	65
62	eat	63
63	in there	63
64	not	62
65	right	62
66	these	62
67	can't	61
68	i wanna	61
69	okay	61
70	sit	61
71	there's	58
72	boy	57
73	he	57
74	three	57
75	book	56
76	look at	56
77	i can't	55
78	take	55
79	hi	54
80	more	54
81	got	53
82	jenko	53
83	read	53
84	who's	53
85	all	52
86	lie	52
87	for	51
88	four	51
89	sleep	51
90	some	51

CHILDES – Naomi (24-35 months) Adjusted (5+)

1	i	224
2	you	209
3	yeah	203
4	uh	198
5	it	173
6	no	165
7	and	153
8	the	149
9	what's that	143
10	that	137
11	a	132
12	my	129
13	want	115
14	on	107
15	this	101
16	me	100
17	is	96
18	don't	95
19	here	92
20	nomi	90
21	i'm	82
22	georgie	81
23	mommy	80
24	to	73
25	daddy	72
26	oh	72
27	up	71
28	yup	71
29	in	67
30	go	66

31	what's	66
32	there	65
33	at	64
34	it's	61
35	okay	61
36	what	56
37	baby	52
38	i want	52
39	one	52
40	this is	52
41	doing	50
42	see	49
43	going	45
44	he	45
45	in the	44
46	i can't	43
47	jenko	42
48	again	41
49	get	41
50	her	41
51	that's	41
52	off	40
53	down	39
54	in there	39
55	too	38
56	sleep	37
57	where's	37
58	who's that	37
59	i wanna	36
60	eyes	34

61	for	34
62	i don't know	33
63	smoke	33
64	was	33
65	yes	33
66	can	31
67	hi	31
68	fish	30
69	have	30
70	now	30
71	all	29
72	another one	29
73	georgie's	29
74	your	29
75	agra	28
76	and a	28
77	boy	28
78	i need	28
79	i want it	28
80	one two three	28
81	uhhuh	28
82	do	27
83	doggie	27
84	get up	27
85	got	27
86	um	27
87	we	27
88	gonna	26
89	him	26
90	in here	26

Michigan Corpus Linguistics

Future developments

- experiment with different thresholds (frequency and statistical)
- incorporate dispersion
- explore more rigorous methods:
  - Mason (2005, 2007) – Automatic Extraction of multi-word units
  - Gries *et al*. - Lexical Gravity Counts (extension of Daudaravicius & Marcinkevibiene 2004)

Michigan Corpus Linguistics

Summary

- Adjusted frequency list is a simple index-based method of producing frequency lists where status of clusters/n-grams as ‘single choice items’ is reflected in frequency of all smaller items
- Initial applications suggest it is able to both reduce number of types in list to be examined and highlights chunks of potential value for both corpus analysis and pedagogy

Hmm… who's been messing with my frequency list?