The adjusted frequency list: A method to produce cluster-sensitive frequency lists

Matthew Brook O'Donnell
University of Michigan

1 Introduction

The suggestion that language learners acquire and make use of multi-word chunks without either breaking them apart or building them up from individual words is well established in psycholinguistic (Pawley and Syder 1983; Ellis 1996, 2003) and corpus linguistic (Sinclair 1991; Stubbs 2002; Meunier and Granger 2008) circles. It is now even discussed in the popular press, as evidenced by a recent edition of the *New York Times* column *On Language* (Zimmer 2010). Frequency lists of items of various lengths are important in both computational and applied linguistics. They are also valuable for measuring the idiomatic/formulaic nature of text (Erman and Warren 2000; Sinclair and Mauranen 2006; Wray 2008). However, many of our computational tools and methods still focus on individual words as the foundational units of analysis.[1] The method proposed here is designed to address this issue.

In discussing the role of chunks in core vocabulary, particularly as it relates to language learners and language teaching, O'Keeffe, McCarthy and Carter highlight the fact that "many chunks are as frequent as or more frequent than the single-word items which appear in the core vocabulary" (2006: 46). Using the CANCODE corpus they found that only 33 single word items appear more frequently than the most common two-word chunk *you know*. Two of those 33 single-word items will be *you* and *know*, and for the latter it seems likely, as O'Keeffe *et al*. suggest, that its high ranking will be due in large part being a part of the highly frequent chunk in spoken English. Such observations highlight the importance of considering the role of chunks of two or more words in the description and teaching of vocabulary (also Nattinger and DeCarrico 1992).

A common methodological step in a corpus linguistic analysis is the extraction of frequency lists of various size chunks (variously called clusters, lexical bundles or n-grams). Most software packages facilitate the creation of such lists, making it possible to compare units of different length. However, each size unit is (necessarily) counted on its own terms without reference to larger units of which they may be a part. For example every instance of *know* is counted individually even if each one of them is preceded by *you*, thus *you know* and *know* have the same frequency. This issue has been discussed with reference to larger units where collecting 3-, 4- and 5-grams together will result in very similar and often identical counts for *at the end*, *the end of*, *at the end of*, *the end of the*, *at the end of the*, *the end of the day* and so on.

The concept of the adjusted frequency list proposed here adjusts the frequency of items of various lengths when they are part of a larger unit that occurs at or above a given frequency or statistical threshold. That is, if *you know* occurs 15 times in a corpus and *know* 20 times, then the frequency of *know* will be adjusted from 20 down to five. The method outlined is 'cluster sensitive' because it boosts the rank of larger word sequences and builds on the notion that if such chunks are single choice items for speakers they should be counted as single items and their internal constituents left uncounted.

Section 2 provides a motivating example for the new procedure of counting words and n-grams which is described in Section 3. The next section describes three potential algorithms to implement the adjusted frequency list procedure. The second (using an index) and third (a two pass process) options are the better approaches and these are used in the case studies reported in Section 6. Two components of the BNC Baby corpus are examined by producing lists of 1- to 5-grams before and after the application of the adjusted frequency list procedure.

2 First interlude: How does a corpus linguist tell a bedtime story?

(1) Once upon a time, there was a little girl named Goldilocks. She went for a walk in the forest. Pretty soon, she came upon a house. She knocked and, when no one answered, she walked right in…

Most readers will be familiar with how this text continues and recognize it as the story of "Goldilocks and the Three Bears" (see Text 1 in the Appendix for full text). How might a typical corpus linguist begin to 'read' (analyze) this particular text? Most likely he or she would begin by generating a word frequency list such as the one in Table 1. As is typical of just about any sample of English, the most frequent types are function words: *the*, *she*, *in*, *and*. These are followed by content words that give some key to who and what the story is about: *chair*, *porridge*, *bear*, *Goldilocks*. From this, therefore, we might answer that a corpus linguist would read this text one word at a time!

Table 1: Frequency list for Top 60 words from Text 1

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*

However, a frequency list of single word items only tells part of the story. Like many stories written for and told to children, "Goldilocks and the Three Bears" contains a certain degree of repetition of phrases, for example, *the three bears*, *someone's been eating my porridge*, *someone's been sleeping in my bed*, *someone's been sitting in my chair*, *growled the papa bear*, *said the mama bear*, *cried the baby bear*. The way to capture these kinds of chunks is to generate a frequency list of n-grams or clusters. While this is often done producing different lists for different values of *n*, it can be valuable to produce a single list covering a range of *n* values. This allows for the kind of comparison between single words and larger chunks alluded to in the quote above from O'Keeffe *et al.* (2006). Table 2 shows such a combined list of 1-, 2- and 3-grams for our bedtime story.[2] The top ten items in the list are still single words but 22 (37%) of the top 60 types are now clusters of two or three words. This suggests the importance of clusters in this text. Now we might want to answer that a corpus linguist would read the text in words AND chunks at the same time.

Table 2: Frequency list of Top 60 1-, 2- and 3-grams from Text 1

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

But again this answer is not without some limitations. Notice how the words *been* and *someone's* have the same frequency (9 occurrences) individually as the bigram *someone's been*. Similarly *eating*, *eating my* and *eating my porridge* all have a frequency of three and likewise *bears*, *the three*, *three bears* and *the three bears*. There are six occurrences of both *so* and *so she* and three of both *baby* and *baby bear*. In each of these cases the largest n-gram accounts for all the occurrences of the smaller n-grams and single words. This raises the question of whether the smaller units should really be included in the frequency list or not. In other instances most but not all of the occurrences of a word can be accounted for by a larger cluster. For example, *into* occurs four times in Table 2 and *into* *the* three times, leaving just one instance of *into* not accounted for by the bigram. The same goes for *been sleeping in* (3 occurrences) and *sleeping* (4 occurrences). In these instances the individual words should certainly remain in the frequency list but their rank appears to be inflated because of the larger cluster.

So is our intrepid corpus linguist perhaps over reading (analyzing) the individual words in the story? How might this issue be addressed?

3 A new concept for frequency counts: The adjusted frequency list

On the wall of my office I have a Dr Seuss ABC poster similar to those often found in a child's nursery or toddler's bedroom. It reads: *A is for Alligator*, *B is for Ball*, *C is for Cat*, and so on. Consider the following 'text' (Text 2), which is 14 tokens long constructed using the first five types:

(2) *Alligator Ball Cat Alligator Ball Cat Alligator Ball Duck Alligator Elephant Alligator Ball Cat*

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

Table 3: Frequency lists of all 1-, 2- and 3-grams from Text 2

Words (1-grams)		2-grams	
Alligator	5	*Alligator Ball*	4
Ball	4	*Ball Cat*	3
Cat	3	*Cat Alligator*	2
Duck	1	*Alligator Elephant*	1
Elephant	1	*Ball Duck*	1
		Duck Alligator	1
		Elephant Alligator	1

3-grams	
Alligator Ball Cat	3
Ball Cat Alligator	2
Cat Alligator Ball	2
Alligator Ball Duck	1
Alligator Elephant Alligator	1
Ball Duck Alligator	1
Duck Alligator Elephant	1
Elephant Alligator Ball	1

If these three lists are merged (again on the basis of frequency and then alphabetically) the list in Table 4 results. For this text the most frequent bigram (*Alligator Ball*) shares the same rank as the second most frequent single item (*Ball*). Similarly, *Alligator Ball Cat*, the most frequent trigram has the same frequency as the second most frequent bigram (*Ball Cat*) and third most frequent single item (*Cat*). From a vocabulary analysis perspective this reinforces the point made by O'Keeffe *et al*. (2006) regarding the value of including clusters in banded frequency lists. N-gram lists are built using a moving window of one word at a time through the text and counting units of length *n*, e.g. *Alligator Ball*, *Ball Cat*, *Cat Alligator* (with *n*=2). This means that aside from the first and last word of a text when collecting units of length *n*, each word is counted *n* times.

Table 4: Combined Frequency list of all 1-, 2- and 3-grams from Text 2

Alligator	5
Alligator Ball	4
Ball	4
Alligator Ball Cat	3
Ball Cat	3
Cat	3
Ball Cat Alligator	2
Cat Alligator	2
Cat Alligator Ball	2
Alligator Ball Duck	1
Alligator Elephant	1
Alligator Elephant Alligator	1
Ball Duck	1
Ball Duck Alligator	1
Duck	1
Duck Alligator	1
Duck Alligator Elephant	1
Elephant	1
Elephant Alligator	1
Elephant Alligator Ball	1

One of the uses of an n-gram list is to discover recurring units that might be formulaic or idiomatic and function as a single choice for the language user (cf. the 'idiom choice principle', Sinclair 1991; Erman and Warren 2000). Setting a threshold for recurrence balances the over counting of the moving window procedure and also serves as a crude measure of formulaicity.

In order to simplify things, consider for a moment a frequency list of Text 2 with all the single words and just the bigrams occurring at least three times (see Table 5).

Table 5: Combined Frequency list of all words and the 2-grams with frequency>2 in Text 2

Alligator	5
Alligator Ball	4
Ball	4
Ball Cat	3
Cat	3
Duck	1
Elephant	1

What this list suggests is that the bigrams *Alligator Ball* and *Ball Cat* are actually single choice units. Ignoring the fact that there is overlap between the units (*Alligator Ball* always overlaps with *Ball Cat*) the text becomes:

(2b) *Alligator Ball Cat Alligator Ball Cat Alligator Ball*
Duck Alligator Elephant Alligator Ball Cat

(where *Alligator Ball* and Ball Cat indicate single units). With the text viewed in this manner the resulting frequency list, shown in Table 6, contains five types and ten tokens:

Table 6: Adjusted Frequency list of all words and the 2-grams with frequency>2 in Text 2b

Alligator Ball	4
Ball Cat	3
Alligator	1
Duck	1
Elephant	1

Notice that single items *Ball* and *Cat* have disappeared from the list because they do not appear independently of the clusters *Alligator Ball* and *Ball Cat*. The count for *Alligator* is reduced from five to just a single occurrence because of the four instances of *Alligator Ball*. I propose the term 'adjusted frequency list' for a frequency list that has undergone this kind of adjustment.

Now what happens if we include trigrams into consideration while keeping the same threshold of three or more occurrences for n-grams. This adds only one item, *Alligator Ball Cat*, to the unadjusted frequency list (see Table 7):

Table 7: Combined Frequency list of all words and the 2- and 3-grams with frequency>2 in Text 2

Alligator	5
Alligator Ball	4
Ball	4
Alligator Ball Cat	3
Ball Cat	3
Cat	3
Duck	1
Elephant	1

Now applying the same adjustment procedure in which longer units (*Alligator Ball Cat*) should take precedence over their component parts (*Alligator Ball*, *Ball Cat*, *Alligator*, *Ball* and *Cat*), the text now consists of five types and seven tokens.

(2c) *Alligator Ball Cat* *Alligator Ball Cat* *Alligator Ball*
Duck Alligator Elephant Alligator Ball Cat

Table 8 contains the adjusted frequency list for Text 2 using a frequency threshold of 3 occurrences for n-grams (with $n>1$). As before single items *Ball* and *Cat* have disappeared and *Alligator* is reduced to a single occurrence. The bigram *Ball Cat* has been removed because it does not occur independently of the trigram *Alligator Ball Cat*, and the four occurrences of *Alligator Ball* have been reduced to the single instance where the bigram is not followed by *Cat*.

Table 8: Adjusted Frequency list of all words and the 2- and 3-grams with frequency>2 in Text 2c

Alligator Ball Cat	3
Alligator	1
Alligator Ball	1
Duck	1
Elephant	1

Although only a toy example, it should be sufficient to illustrate the notion of the adjusted frequency list. There are a number of parameters, particularly the thresholds to use at various values of *n* and the maximum value of *n*, that can be tuned and will result in different outputs. But the key characteristic of the procedure is that it is sensitive to the use of clusters as (potentially) single lexical choices.

4 Algorithms for the adjusted frequency list procedure

The previous section provided an overview of the adjusted frequency list procedure without any suggestion of how it might be implemented. This section presents three possible algorithms in some detail. It is not necessary to follow through the details of each algorithm and this whole section can be skimmed over without losing the overall concept of the adjusted frequency list.

4.1 Simple non-indexed algorithm

The first and simplest approach is applied just to the frequency list of 1-, 2-, … N_{max}-grams. Given a text or set of texts an adjusted frequency list is constructed in the following manner.

1. Construct frequency lists (or a single combined list) for all items length 1 to N_{max} 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_{max} that occur with frequency less than desired threshold adopted for formula/unit status.
3. For each remaining n-gram with frequency *f* (in descending order by length, i.e. N_{max} to 2) derive each of its component sub-items.

   So for the trigram *Alligator Ball Cat* there are bigrams *Alligator Ball* and *Ball Cat* and three single items *Alligator*, *Ball* and *Cat*.
4. Reduce the frequency of each of these sub-items by *f* (unless frequency=0).

In essence this algorithm groups all items in a combined frequency list into a tree (or directed graph) with larger n-grams higher up the tree linked to smaller n-grams that are component parts of the larger n-gram. But the trigram *Alligator Ball Cat* will link not only to bigrams *Alligator Ball* and *Ball Cat* but also to each of individual words *Alligator*, *Ball* and *Cat*.

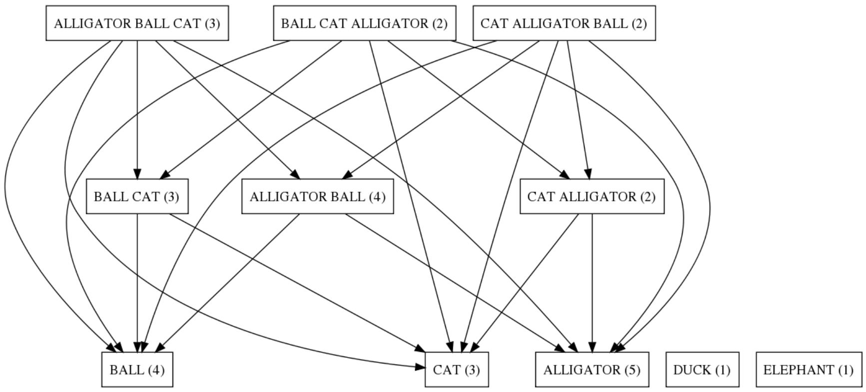
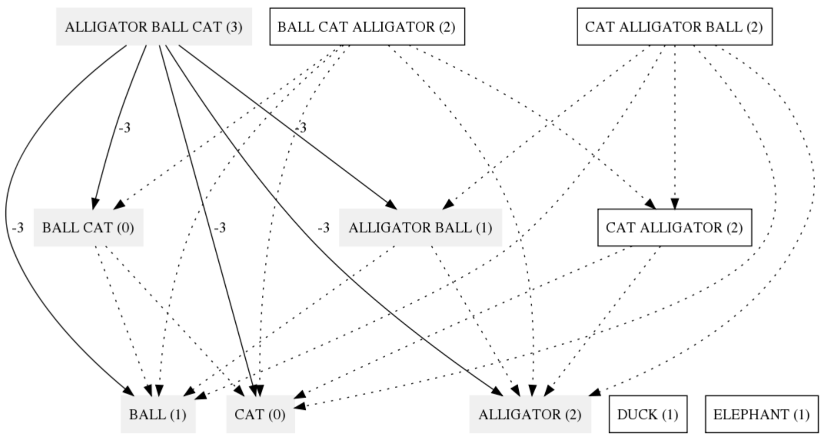


Figure 1: Links between n-grams in simple non-indexed algorithm

Figure 1 illustrates these connections for the 2- and 3-grams in Text 2 with a frequency of 2 or more extracted from Table 4. Figure 2 shows the first two iterations of the algorithm at Step 3 for the n-grams *Alligator Ball Cat* and *Ball Cat Alligator*.

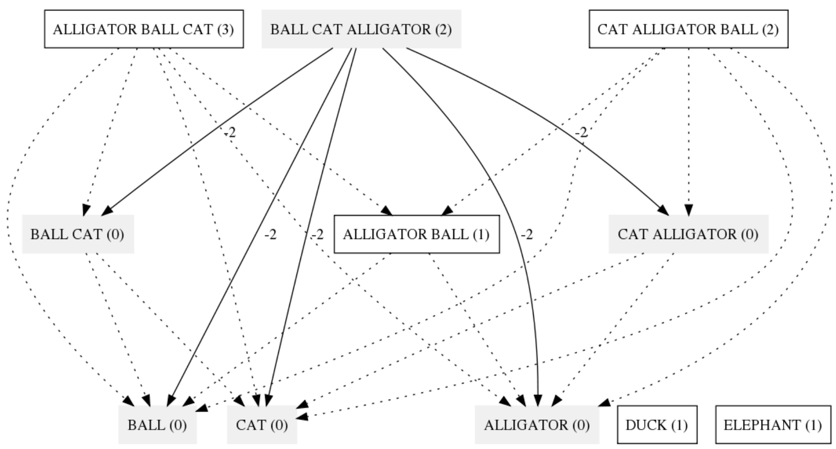
Reduction process (Step 3)



Resulting frequency list
3 *Alligator Ball Cat*
2 *Alligator*
2 *Ball Cat Alligator*
2 *Cat Alligator*
2 *Cat Alligator Ball*
1 *Alligator Ball*
1 *Ball*
1 *Duck*
1 *Elephant*
0 *Ball Cat*
0 *Cat*

Figure 2a: Applying the simple non-indexed algorithm to n-grams in Text 2.

Reduction process (Step 3)



Resulting frequency list
3 *Alligator Ball Cat*
2 *Ball Cat Alligator*
2 *Cat Alligator Ball*
1 *Alligator Ball*
1 *Duck*
1 *Elephant*
0 *Alligator*
0 *Ball*
0 *Ball Cat*
0 *Cat*
0 *Cat Alligator*

Figure 2b: Applying the simple non-indexed algorithm to n-grams in Text 2

The problem with this simple method is that it is likely to be too productive in the final step. That is, given trigrams [*Alligator* $Ball_i$ *Cat*], [$Ball_i$ *Cat Duck*], [*Alligator* $Ball_j$ *Cat*] and [$Ball_j$ *Cat Duck*], generated from the string *Alligator*

$Ball_i$ Cat Duck Alligator $Ball_j$ Cat Duck, the counts for both *$Ball_i$* and *$Ball_j$* will be reduced twice. This is because the procedure has no knowledge of which particular *Ball* is being referenced. This is further illustrated in Figure 2, where after applying reductions to just two trigrams (*Alligator Ball Cat* and *Ball Cat Alligator*) the count for the single item *Alligator* has been reduced to zero. We know from Table 8 that the final count for *Alligator* should actually be one after applying the full procedure.

4.2 Indexed algorithm

To address the limitation of the simplest possible algorithm two further algorithms are presented. The first builds an index from the corpus and then can selectively reduce counts for smaller values of *n* as it reduces a specific n-gram. Given a text or set of texts an adjusted frequency list is constructed in the following manner:

1. Construct indexed frequency lists for all items length 1 to N_{max}, so that each instance of an item is recorded with reference to 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_{max} 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_{max} to 2) derive each of its component sub-items, recording the start and end positions for each occurrence of the n-gram.
4. For each of the sub-items identified in Step 3, scan their index records for an occurrence that falls within the position range of the larger n-gram and remove record.

So given Text 3:

(3) *$Alligator_1$ $Ball_2$ Cat_3 $Alligator_4$ $Ball_5$ Cat_6 $Alligator_7$ $Ball_8$ Cat_9*

where the subscripts indicate word (or position) offset, the following index entries would result from Step 1. (Each instance of an item is recorded with the form startOffset:endOffset).

Alligator Ball Cat	[1:3, 4:6, 7:9]	3
Ball Cat Alligator	[2:4, 5:7]	2
Cat Alligator Ball	[3:5, 6:8]	2
Alligator Ball	[1:2, 4:5, 7:8]	3
Ball Cat	[3:4, 5:6, 8:9]	3
Cat Alligator	[3:4, 6:7]	2
Alligator	[1:1, 4:4, 7:7]	3
Ball	[2:2, 5:5, 8:8]	3
Cat	[3:3, 6:6, 9:9]	3

After the application of Step 2 with a threshold of 3, the index would be:

Alligator Ball Cat	[1:3, 4:6, 7:9]	3
Alligator Ball	[1:2, 4:5, 7:8]	3
Ball Cat	[3:4, 5:6, 8:9]	3
Alligator	[1:1, 4:4, 7:7]	3
Ball	[2:2, 5:5, 8:8]	3
Cat	[3:3, 6:6, 9:9]	3

Step 3 would begin with the trigram *Alligator Ball Cat* and derive the sub-items *Alligator Ball*, *Ball Cat*, *Alligator*, *Ball* and *Cat*. For each entry in the index for *Alligator Ball Cat* the entries for these sub-items is scanned for entries that fall within the start and end offsets. Matching entries are deleted, as follows:

Alligator Ball Cat	[**1:3,** 4:6, 7:9]	3
Alligator Ball	[~~1:2~~, 4:5, 7:8]	2
Ball Cat	[~~2:3~~, 5:6, 8:9]	2
Alligator	[~~1:1~~, 4:4, 7:7]	2
Ball	[~~2:2~~, 5:5, 8:8]	2
Cat	[~~3:3~~, 6:6, 9:9]	2

Alligator Ball Cat	[1:3, **4:6**, 7:9]	3
Alligator Ball	[~~4:5~~, 7:8]	1
Ball Cat	[~~5:6~~, 8:9]	1
Alligator	[~~4:4~~, 7:7]	1
Ball	[~~5:5~~, 8:8]	1
Cat	[~~6:6~~, 9:9]	1

Alligator Ball Cat	[1:3, 4:6, **7:9**]	3
Alligator Ball	[~~7:8~~]	0
Ball Cat	[~~8:9~~]	0
Alligator	[~~7:7~~]	0
Ball	[~~8:8~~]	0
Cat	[~~9:9~~]	0

In this extreme case the adjusted frequency list contains a single trigram *Alligator Ball Cat* with a frequency of 3. All instances of the 5 sub-items, *Alligator Ball*, *Ball Cat*, *Alligator*, *Ball* and *Cat* occur within these the three instances of the trigram.

4.3 The Serial Cascading Algorithm

An alternative approach that does not need an index but avoids the problems of the non-indexed approach discussed in Section 4.1 has been suggested by Catherine Smith (p.c.). It takes two passes over the texts in a corpus. The first pass constructs the relevant n-gram lists and the second pass counts n-grams according to a largest *n* first cascade:

Pass #1

1. Construct frequency lists (or a single combined list) for all items length 2 to N_{max} 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_{max} that occur less than desired threshold used for formula/unit status.

Pass #2

3. Initialize:
   a. adjusted_list = {}
   b. p = 1
   c. $last_i = \varnothing$
4. Step through using a moving window of one token steps using position counter p.
5. Select n-gram$_{candidate,}$ an n-gram of N_{max}
6. Check whether n-gram$_{candidate}$ is found in lists constructed in PASS #1
   a. If yes and $p + N_{max} -1 > last_i$ add one to the count for ngram$_{candidate}$ in the adjusted list, set $last_i$ to $p + N_{max}$ and return to Step 4
   b. else reduce N_{max} by 1
      i. If $N_{max} > 1$
         1. If $p + N_{max} – 1 > last_i$ return to Step 5
         2. else reset N_{max} and return to Step 4
      ii. else if $p > last_i$ add one to single word count in the adjusted list and return to Step 4

If this algorithm is applied to Text 2 (used in Section 3) using a frequency threshold of three or greater for 2- and 3-grams the algorithm proceeds as follows:

Pass #1

Collect all 2- and 3-grams occurring three or more times in text.

Alligator Ball	4
Alligator Ball Cat	3
Ball Cat	3

Pass #2

The second part of the algorithm is somewhat complex. Below three snapshots of the process as applied to Text 2 are illustrated with the value of variables at each step shown:

Step	**Location in text** (n-gram$_{candidate}$ in bold)	**Variables**	**Adjusted list**
3	*Alligator Ball Cat* *Alligator Ball Cat* *Alligator Ball Duck* *Alligator Elephant* *Alligator Ball Cat*	p=1 $last_i=\varnothing$	{}

4
5
 	
6
6a
	Alligator Ball Cat
	Alligator Ball Cat
	Alligator Ball Duck
	Alligator Elephant
	Alligator Ball Cat
		p=1
		$last_i=\varnothing$
		$N_{max}=3$
		Alligator Ball Cat on list
		$last_i=3$
			{
			'Alligator Ball Cat': 1
			}

4
5
 	
6
6b
6b.i.
2
	*Alligator **Ball Cat***
	***Alligator** Ball Cat*
	Alligator Ball Duck
	Alligator Elephant
	Alligator Ball Cat
		p=2
		$last_i=3$
		$N_{max} = 3$
		Ball Cat Alligator not on list
		$N_{max} = 2$
		$p + N_{max} -1 = last_i$
			{
			'Alligator Ball Cat': 1
			}

4
5
 	
6
6b
6b.i.
1
	*Alligator Ball **Cat***
	***Alligator Ball** Cat*
	Alligator Ball Duck
	Alligator Elephant
	Alligator Ball Cat
		p=3
		$last_i=3$
		$N_{max}=3$
		Cat Alligator Ball not on list
		$N_{max}= 2$
		$p + N_{max} -1 > last_i$
			{
			'Alligator Ball Cat': 1
			}

5
 	
6
6b
6b.i
6b.i.
2
	*Alligator Ball **Cat***
	***Alligator** Ball Cat*
	Alligator Ball Duck
	Alligator Elephant
	Alligator Ball Cat
		p=3
		$last_i=3$
		$N_{max}=2$
		Cat Alligator not on list
		$N_{max}=1$
		$p + N_{max}–1 = last_i$
		$N_{max}=3$
			{
			'Alligator Ball Cat': 1
			}

4
5
 	
6
6a
	Alligator Ball Cat
	Alligator Ball Cat
	Alligator Ball Duck
	Alligator Elephant
	Alligator Ball Cat
		p=4
		$last_i=3$
		$N_{max}=3$
		Alligator Ball Cat on list
		$last_i=6$
			{
			'Alligator Ball Cat': 1
			}

	(some intervening steps skipped)		

5
 	
6
6a
	Alligator Ball Cat
	Alligator Ball Cat
	***Alligator Ball** Duck*
	Alligator Elephant
	Alligator Ball Cat
		p=7
		$last_i$=6
		N_{max}=2
		Alligator Ball on list
		$p + 2 -1 > last_i$
		$last_i$=8
		N_{max}=3
			{
			‘Alligator Ball Cat’: 2,
			‘Alligator Ball’: 1
			}

4
5
 	
6
6b
6b.i.
1
	Alligator Ball Cat
	Alligator Ball Cat
	*Alligator **Ball Duck***
	***Alligator** Elephant*
	Alligator Ball Cat
		p=8
		$last_i$=8
		N_{max}=3
		Ball Duck Alligator not on list
		N_{max}=2
		$p + 2 -1 > last_i$
			{
			‘Alligator Ball Cat’: 2,
			‘Alligator Ball’: 1
			}

5
 	
6
6b
6b.i
6b.i.
2
	Alligator Ball Cat
	Alligator Ball Cat
	*Alligator **Ball Duck***
	Alligator Elephant
	Alligator Ball Cat
		p=8
		$last_i$=8
		N_{max} = 2
		Ball Duck not on list
		N_{max}=1
		$p + 1 -1 = last_i$
		N_{max}=3
			{
			‘Alligator Ball Cat’: 2,
			‘Alligator Ball’: 1
			}

	(some intervening steps skipped)		

5
 	
6
6b
6b.ii
	Alligator Ball Cat
	Alligator Ball Cat
	Alligator Ball Duck
	*Alligator **Elephant***
	***Alligator** Ball Cat*
		p=11
		$last_i$=10
		N_{max} = 2
		Elephant Alligator not on list
		N_{max}=1
		$p > last_i$
			{
			‘Alligator Ball Cat’: 2,
			‘Alligator Ball’: 1,
			‘Duck’: 1,
			‘Alligator’: 1,
			‘Elephant’: 1
			}

4
5
 	
6
6a
	Alligator Ball Cat
	Alligator Ball Cat
	Alligator Ball Duck
	Alligator Elephant
	Alligator Ball Cat
		p=12
		$last_i$=10
		N_{max} = 3
		Alligator Ball Cat on list
		$last_i$=11
			{
			‘Alligator Ball Cat’: 3,
			‘Alligator Ball’: 1,
			‘Duck’: 1,
			‘Alligator’: 1,
			‘Elephant’: 1
			}

The resulting output is:

Alligator Ball Cat	3
Alligator Ball	1
Alligator	1
Duck	1
Elephant	1

Future work is needed to implement and test these (and other algorithms) to gauge the applications for which each is best suited. The use of an index requires more in terms of computation resources but does allow for comparative concordancing of unadjusted and adjusted items. The Serial Cascading Algorithm is more lightweight and could potentially scale to a distributed/parallel implementation.

5 Second interlude: The well-adjusted bedtime story

We now return to our corpus linguist endeavoring to tell a bedtime story using the state-of-the-art tools of the trade. When we left them back in Section 2, they had begun to come to terms with the highly repetitious and chunky nature of the typical bedtime story and created a combined 1- to 3-gram list (see Table 2). But there were at least two problems with this approach. First, single words still fill the top ranks of the list even though many of them are components of highly frequent chunks of two or three words. And second, a number of frequent bigrams on the list were entirely accounted for by certain trigrams.

Table 9 shows the top of the adjusted frequency list for Text 1 for 1-, 2- and 3-grams using a frequency threshold of three or more occurrences for the inclusion of 2- and 3-grams in the adjustment process. When compared to the unadjusted list in Table 2 notice the marked reduction for *the* from 34 occurrences down to nine. This indicates that 25 instances of *the* are a part of bi- or trigram that occurs three times or more. Likewise the 9 occurrences of *my* (rank 9 in Table 1) are all accounted for by the trigrams: *eating my porridge* (3), *in my chair* (3) and *in my bed* (3). Notice also how the bigrams *been eating*, *been sitting*, *been sleeping*, all with three occurrences in Table 2, no longer occur in the adjusted list. This is because they are fully accounted for by the larger trigrams *been eating my*, *been sitting in* and *been sleeping in*.

Table 9: Adjusted Frequency list of Top 60 1-, 2- and 3-grams from Text 1

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

The five instances of *just* in Table 1 become three of *just right* and two for single-item *just* in Table 9. A side-effect of this grouping is that the two adverbial usages of *just* have been distinguished: i. exactly (*just right*) in lines 1,2 and 4 and ii. temporal in lines 3 and 5.

```
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
```

It is worth calling attention to a couple of points arising from the adjusted list in Table 9 that illustrate the effects of choices made with regards to the largest n-gram (N_{max}) included in the adjustment procedure and also the threshold (or thresholds) chosen for the different values of *n*. Here the procedure was applied at N_{max}=3 with a threshold of 3+ occurrences. As a result there remains some overlapping n-grams that are actually part of a larger chunk. For instance, *been eating my* and *someone's been eating* both have 3 occurrences in the list in Table 9. These are clearly part of a larger 4-gram *someone's been eating my*. Likewise towards the end of the list we can see overlap between words and bigrams with frequencies below the selected threshold—for example: *ahh* and *ahh this* with 2 occurrences.[3]

These minor caveats aside, our corpus linguist now has a tool that provides a more realistic picture of the interaction of chunks and single words in the Goldilocks text. And the example of *just* demonstrates the potential of improved efficiency in a KWIC analysis, which as everyone knows is both the next act in the story and another story all by itself (see O'Donnell 2008).

6 *Looking at some larger corpora*

The final two examples apply the adjusted frequency list method to two of the categories in the BNC Baby sample corpus. These two sections are the 1 million word demographically sampled spoken component (30 texts) and the 1 million word sub-corpus of academic texts (also 30 texts).

6.1 BNC Baby Demographic section

The list in Table 10 contains the top 150 1-, 2- and 3-grams from the Demographic section of the corpus with no adjustment.

Table 10: Top 150 combined 1-, 2- and 3-grams according to type frequency in BNC Baby Demographic section

Rank	***Item***	***Freq.***	***Rank***	***Item***	***Freq.***	***Rank***	***Item***	***Freq.***
1	*i*	30371	51	*up*	4056	101	***it was***	1901
2	*you*	29688	52	*with*	3833	102	*very*	1878
3	*the*	27698	53	*erm*	3813	103	*can't*	1868
4	*it*	21834	54	*them*	3670	104	*five*	1855
5	*and*	19845	55	*at*	3662	105	*four*	1820
6	*a*	19600	56	*are*	3652	106	***on the***	1789
7	*to*	17180	57	*me*	3607	107	*been*	1770
8	*that*	14722	58	***you know***	3605	108	*bit*	1715

9	*yeah*	14303	59	*said*	3563	109	*alright*	1703
10	*oh*	10398	60	*two*	3528	110	*would*	1657
11	*in*	10133	61	*your*	3448	111	*him*	1655
12	*no*	9804	62	*out*	3168	112	*they're*	1653
13	*of*	9799	63	*i'm*	3153	113	*were*	1625
14	*it's*	8534	64	*see*	3143	114	***i know***	1623
15	*well*	8478	65	*now*	3081	115	*back*	1590
16	*what*	8171	66	*or*	3005	116	*time*	1580
17	*on*	7951	67	*did*	2911	117	*only*	1578
18	*is*	7816	68	***i don't***	2878	118	*you've*	1569
19	*have*	7802	69	*when*	2855	119	*off*	1555
20	*know*	7659	70	*had*	2829	120	*why*	1535
21	*one*	7488	71	*about*	2825	121	*something*	1510
22	*do*	7280	72	*want*	2823	122	*where*	1508
23	*was*	7133	73	*cos*	2796	123	***don't know***	1495
24	*got*	6842	74	*as*	2750	124	*could*	1486
25	*we*	6686	75	*mean*	2716	125	*she's*	1453
26	*he*	6618	76	***in the***	2662	126	*will*	1444
27	*don't*	6477	77	*my*	2504	127	*because*	1442
28	*they*	6475	78	*going*	2377	128	***have to***	1431
29	*but*	6178	79	***i mean***	2364	129	***you can***	1398
30	*so*	6148	80	*i've*	2327	130	***is it***	1390
31	*there*	6125	81	*put*	2303	131	*ah*	1380
32	*that's*	5957	82	***i think***	2286	132	*from*	1362
33	*for*	5673	83	*here*	2270	133	*his*	1358
34	*mm*	5662	84	*really*	2238	134	***if you***	1315
35	*not*	5270	85	*i'll*	2214	135	*nice*	1314
36	*go*	4941	86	*he's*	2212	136	*an*	1296
37	*be*	4869	87	***do you***	2196	137	*isn't*	1283
38	*this*	4781	88	*come*	2185	138	*mum*	1282
39	*get*	4772	89	*three*	2181	139	*what's*	1278
40	*like*	4744	90	*down*	2147	140	*thought*	1261
41	*just*	4696	91	*look*	2099	141	*any*	1254

42	*she*	4683	92	*didn't*	2074	142	*little*	1241
43	*all*	4459	93	*how*	2063	143	***of the***	1233
44	*er*	4441	94	*good*	2050	144	***and then***	1226
45	*yes*	4432	95	*you're*	2044	145	*more*	1220
46	*then*	4369	96	*there's*	2040	146	*haven't*	1212
47	*right*	4252	97	*gonna*	2007	147	***i don't know***	1204
48	*if*	4234	98	*her*	1983	148	***and i***	1192
49	*think*	4159	99	*some*	1950	149	*hundred*	1191
50	*can*	4148	100	*say*	1923	150	*much*	1180

There are 18 (12%) 2- or 3-grams among the top 150 items. This confirms the observation by O'Keeffe *et al*. (2006: 46) from their analysis of CANCODE concerning the high frequency of many chunks in spoken corpora. These items are marked in bold in Table 10. 12 of the 18 are in the third column of the list and thereby have a rank of 100 or greater. The first is *you know* at rank 58 with 3605 occurrences. The component words of this bigram are found at rank 2 (*you* 29688 occurrences) and rank 20 (*know* 7659 occurrences). The sole trigram in the top 150 items is *i don't know* with 1204 occurrences at rank 147. The unadjusted list should be compared with the adjusted frequency list in Table 11, where the procedure has been applied using a threshold value of five for both bigrams and trigrams for inclusion in the adjustment process. 43 (28.7%) of the top 150 items in the adjusted list are bi- or tri-gram items (marked in bold). The most frequent trigram in the adjusted list is *i don't know* (a move from rank 147 to rank 7). The three component words have experienced significant reduction: *i* (30371 [rank 1] → 660 [rank 17]), *don't* (6477 [rank 27] → 188 [rank 167]) and *know* (7659 [rank 20] → 51 [rank 1403]).

Table 11: Top 150 combined 1-, 2- and 3-grams in BNC Baby Demographic section after adjustment (using threshold for 2- and 3-grams of 5+ occs)

Rank	***Item***	***Freq.***	***Rank***	***Item***	***Freq.***	***Rank***	***Item***	***Freq.***
1	*yeah*	6877	51	*my*	398	101	***a bit of***	262
2	*mm*	3866	52	***a lot of***	395	102	***have a look***	258
3	*no*	3026	53	*that's*	392	103	***do you think***	255
4	*oh*	2002	54	*me*	380	104	*by*	254
5	*and*	1787	55	***isn't it***	373	105	***oh dear***	251
6	*yes*	1671	56	*this*	373	106	***on the***	249
7	***i don't know***	1204	57	*ha*	365	107	*four*	248
8	*the*	1103	58	*up*	364	108	*she's*	248
9	*what*	1100	59	*like*	361	109	*she*	246
10	*right*	887	60	***what do you***	356	110	*they*	246
11	*er*	808	61	***and the***	349	111	*down*	245
12	*erm*	800	62	*here*	338	112	***have you got***	244
13	*a*	754	63	***no no***	335	113	*are*	240
14	*in*	743	64	*please*	325	114	***three four five***	238
15	*that*	732	65	*who*	325	115	*first*	237
16	*well*	703	66	***i know***	324	116	*aye*	236
17	*i*	660	67	*just*	324	117	*good*	234
18	*or*	654	68	***that's right***	323	118	*aha*	232
19	*ah*	648	69	*anyway*	321	119	*him*	232
20	*it's*	630	70	*again*	319	120	***i think***	232
21	*then*	620	71	*out*	314	121	*where*	232
22	*it*	617	72	*today*	310	122	***for the***	231
23	*of*	591	73	*innit*	307	123	***of the***	228
24	*now*	568	74	***you have to***	307	124	*sorry*	228
25	*there*	564	75	*two*	306	125	***you've got to***	226
26	*you*	563	76	*eh*	304	126	*bloody*	224
27	*so*	546	77	*look*	304	127	*any*	223
28	*to*	544	78	*why*	304	128	*our*	223
29	***do you want***	543	79	*though*	300	129	*they're*	223
30	*for*	533	80	***yeah yeah***	297	130	*which*	223
31	***one two three***	520	81	*from*	296	131	***you know i***	221

32	***mm mm***	518	82	*he*	295	132	*pardon*	219
33	***i don't think***	517	83	*not*	295	133	*yep*	219
34	*ooh*	500	84	***as well***	294	134	*it's a*	217
35	*one*	491	85	*at*	294	135	***oh yes***	217
36	*is*	485	86	*them*	293	136	*their*	217
37	*mum*	482	87	*hello*	289	137	*you're*	217
38	***oh yeah***	481	88	*his*	286	138	***come on***	216
39	*really*	481	89	*was*	281	139	*daddy*	216
40	*but*	471	90	***you want to***	281	140	*probably*	216
41	*on*	470	91	*i'm*	278	141	*bye*	212
42	***two three four***	449	92	*her*	277	142	*some*	212
43	***is it***	447	93	***i mean i***	276	143	*these*	212
44	*with*	440	94	*as*	275	144	***with the***	212
45	***you know***	431	95	*off*	274	145	*oh no*	211
46	*your*	430	96	*dad*	273	146	*an*	210
47	*alright*	423	97	*actually*	272	147	***thank you***	207
48	*mhm*	411	98	***in the***	271	148	*we*	206
49	*mummy*	409	99	*he's*	267	149	***you know what***	206
50	*okay*	406	100	***no no no***	267	150	***that's it***	205

There is strong support, particularly in the case of *know*, for the claim that a standard (unadjusted) frequency list considerably inflates the frequency of single words that belong to larger chunks. Aside from *i don't know*, notice *you know* (rank 45), *i know* (rank 66), *you know I* (rank 131) and *you know what* (rank 149) as chunks containing *know*. In fact in the adjusted list there are 45 bi- and trigrams containing *know* with a higher rank than the single word item *know*. None of which, of course, were found above *know* in the unadjusted list.

Another interesting observation concerning differences between the unadjusted (Table 10) and adjusted (Table 11) frequency lists from BNC Baby Demographic is the rank reduction of many of the function words that routinely top any English frequency list. While the top ranking of personal pronouns *i* and *you* in the unadjusted list, above *the*, are an indication of spoken language, the top of the list is still quite generic. After adjustment, however, most of these items have dropped significantly in rank because of their participation in frequent chunks. The top of the adjusted list is now much more distinctly speech-like: *yeah*, *mm*, *oh*, *no*, *yes*, *right*. Further, many of the bi- and tri-gram chunks in the adjusted list 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*) or declarative statements (*i don't know*, *i don't think*, *i think it's*, *i want to*). Such observations need more detailed examination along with the application of the procedure to other spoken corpora.

6.2 BNC Baby Academic section

Table 12 shows the top 150 items from the frequency list for the academic section of the BNC Baby combining single words and 2- and 3-grams. 29 (19.3%) of the top 150 items are bi- or trigrams (marked in bold):

Table 12: Top 150 combined 1-, 2- and 3-grams according to type frequency in BNC Baby Academic section

Rank	*Item*	*Freq.*	*Rank*	*Item*	*Freq.*	*Rank*	*Item*	*Freq.*
1	the	70257	51	he	2219	101	***there is***	1071
2	of	44195	52	had	2134	102	social	1066
3	and	26672	53	***that the***	2132	103	work	1061
4	to	26245	54	than	2114	104	because	1053
5	in	25169	55	***on the***	2096	105	both	1044
6	a	23047	56	all	2075	106	***is not***	1042
7	is	17401	57	***of a***	2075	107	do	1035
8	that	12449	58	so	2019	108	***may be***	1027
9	***of the***	10454	59	his	2000	109	***as the***	1025
10	be	9364	60	i	1839	110	case	994
11	for	9303	61	its	1835	111	***at the***	984
12	as	9243	62	also	1784	112	now	983
13	it	8209	63	only	1765	113	history	980
14	are	7551	64	would	1759	114	3	978
15	by	7246	65	when	1730	115	used	970
16	this	7052	66	between	1728	116	even	969
17	with	6991	67	what	1709	117	being	964
18	***in the***	6244	68	***for the***	1668	118	***has been***	960
19	which	6026	69	no	1651	119	up	951
20	on	5779	70	***by the***	1642	120	***have been***	949
21	or	5508	71	you	1631	121	form	935

22	not	5192	72	about	1597	122	very	932
23	was	5077	73	2	1507	123	each	921
24	have	4710	74	1	1501	124	species	921
25	from	4603	75	***can be***	1484	125	same	889
26	an	4491	76	where	1478	126	per	886
27	but	3890	77	who	1449	127	new	883
28	at	3612	78	***with the***	1447	128	must	882
29	we	3425	79	then	1431	129	***the same***	875
30	can	3238	80	two	1425	130	***in this***	872
31	***to the***	3225	81	any	1412	131	way	870
32	has	3159	82	***from the***	1407	132	law	862
33	there	3032	83	***in a***	1404	133	general	852
34	they	3001	84	into	1363	134	***this is***	841
35	were	2885	85	***as a***	1360	135	rather	838
36	their	2795	86	most	1354	136	***to a***	835
37	been	2790	87	those	1333	137	how	834
38	***it is***	2757	88	time	1255	138	b	831
39	more	2729	89	***is a***	1248	139	people	809
40	if	2705	90	them	1214	140	particular	808
41	one	2681	91	should	1205	141	much	807
42	formula	2635	92	example	1168	142	number	806
43	may	2619	93	however	1153	143	over	803
44	***and the***	2399	94	many	1146	144	might	800
45	***to be***	2384	95	***is the***	1127	145	within	800
46	such	2295	96	different	1122	146	data	797
47	these	2288	97	first	1115	147	see	791
48	will	2281	98	out	1101	148	given	782
49	other	2259	99	could	1093	149	does	757
50	some	2220	100	use	1074	150	thus	743

The unadjusted list should be compared with the list in Table 13 which is the adjusted frequency list using thresholds for 2- and 3-grams of 10+ and 20+ occurrences respectively. The use of two different thresholds here is simply to demonstrate how the parameters can be varied. Future work is required to ascertain the most appropriate thresholds for different size corpora, different genres and different ranges of *n*. 40 of the top 150 items (26%) in the adjusted list are bi- or tri-grams compared to 29 of the top 150 in the unadjusted list (Table 12). This is a less marked change than the one seen with the Demographic section. Also there are no trigrams in the top 150 items either before or after adjustment.

Table 13: Top 150 combined 1-, 2- and 3-grams in BNC Baby Academic section after adjustment

Rank	***Item***	***Freq.***	***Rank***	***Item***	***Freq.***	***Rank***	***Item***	***Freq.***
1	*and*	12984	51	*when*	747	101	*other*	451
2	*the*	8500	52	*its*	723	102	*b*	446
3	*of*	7536	53	*can*	690	103	*like*	444
4	*in*	5523	54	*into*	689	104	*most*	444
5	*to*	4715	55	*1*	678	105	***for a***	442
6	*a*	4052	56	*if*	676	106	*while*	437
7	*or*	3903	57	*had*	672	107	***and a***	434
8	*for*	3414	58	*2*	668	108	*first*	429
9	***of the***	3225	59	*they*	661	109	*you*	429
10	*is*	3168	60	*we*	645	110	*here*	422
11	*by*	3094	61	*such*	635	111	*be*	421
12	***in the***	2705	62	***to a***	634	112	*without*	419
13	*are*	2667	63	*may*	627	113	*after*	417
14	*as*	2647	64	*then*	616	114	*work*	416
15	*with*	2330	65	***as the***	615	115	*thus*	413
16	*that*	2245	66	***can be***	605	116	*7*	412
17	***and the***	2110	67	*where*	605	117	***which is***	410
18	*was*	2044	68	***it is***	598	118	***but the***	406
19	*this*	1971	69	*only*	596	119	*often*	406
20	*which*	1948	70	***to be***	594	120	***if the***	404
21	*on*	1807	71	*however*	592	121	*c*	401
22	*from*	1719	72	*3*	588	122	*5*	400
23	***to the***	1680	73	***is the***	587	123	***about the***	400
24	*were*	1663	74	*also*	575	124	*10*	397
25	*an*	1445	75	***with a***	574	125	*species*	396

26	*but*	1405	76	***is a***	557	126	*her*	391
27	***of a***	1327	77	*who*	550	127	*very*	390
28	***for the***	1252	78	*any*	543	128	*data*	384
29	***by the***	1170	79	*both*	537	129	*6*	382
30	***on the***	1145	80	*so*	526	130	*many*	382
31	*their*	1113	81	*all*	525	131	*social*	380
32	***from the***	1100	82	*no*	520	132	***should be***	378
33	*it*	1087	83	*them*	520	133	***have been***	376
34	***with the***	1052	84	*4*	513	134	*through*	374
35	***that the***	1015	85	*formula*	509	135	*further*	373
36	*at*	985	86	*being*	505	136	*how*	371
37	***in a***	968	87	*would*	500	137	*see*	370
38	*these*	959	88	***such as***	497	138	*before*	369
39	*has*	953	89	*two*	496	139	***for example***	369
40	*his*	894	90	*not*	495	140	***of an***	368
41	*one*	881	91	*than*	495	141	***will be***	368
42	*some*	875	92	***at the***	487	142	*she*	367
43	*more*	863	93	*over*	484	143	*itself*	366
44	*have*	830	94	*now*	483	144	*should*	366
45	*about*	814	95	***has been***	472	145	*since*	363
46	*he*	805	96	***may be***	472	146	*could*	362
47	*will*	794	97	*p*	467	147	***is not***	361
48	*i*	792	98	*what*	465	148	*time*	361
49	***as a***	785	99	***of this***	455	149	***of their***	360
50	*between*	766	100	***by a***	452	150	*history*	358

Table 14 lists all of these chunks from the unadjusted and adjusted frequency lists with an indication of how their rank has changed after adjustment (↓ means reduction in rank [8 items], ↑ an increase [33 items] and = if it stays the same [1 item]). All of these items are bigrams so they have all dropped in frequency from the unadjusted list because the procedure begun with trigrams. The increase in rank of the majority of the items may be partially responsible for the drop in rank of the following items: *it is*, *to be*, *have been*, *is not*, *there is*, *the same*, *in this*, *this is*. However, each of these is part of a frequent trigram, e.g. *the* *same* > *the same time*, *the same as*, *the same way*; *is not* > *is not a*, *is not the*, *is not to*, *is not surprising*; *there is* > *there is a/an*, *there is no*, *there is little*, *there is some*, *there is also*, *there is evidence*. In the case of *there is*, it is less frequent in the adjusted list (163 occurrences) than two 3-grams: *there is a* (301 occurrences) and *there is no* (245 occurrences).

Table 14: All bigrams in the top 150 items from 1-, 2- and 3-gram lists of BNC Baby Academic showing rank change after adjustment

item	**change of rank?**	**rank before**	**rank after**	**frequency before**	**frequency after**
of the	=	9	9	10454	3299
in the	↑	18	13	6244	2636
and the	↑	44	17	2399	2069
to the	↑	31	23	3225	1625
for the	↑	68	27	1668	1234
of a	↑	57	28	2075	1192
by the	↑	70	29	1642	1161
from the	↑	82	32	1407	1089
on the	↑	55	33	2096	1059
that the	↑	53	34	2132	1028
with the	↑	78	35	1447	1002
in a	↑	83	36	1404	998
as a	↑	85	49	1360	766
to a	↑	136	61	835	640
<u>*it is*</u>	↓	38	62	2757	636
<u>*to be*</u>	↓	45	67	2384	610
as the	↑	109	69	1025	596
can be	↑	75	70	1484	590
with a	↑	173	73	670	560
is a	↑	89	76	1248	540
is the	↑	95	77	1127	540
such as	↑	163	89	696	494
has been	↑	118	90	960	493

may be	↑	108	93	1027	480
at the	↑	111	98	984	455
for a	↑	209	102	568	439
by a	↑	241	103	500	433
and a	↑	280	105	444	431
of this	↑	155	109	719	429
which is	↑	205	115	572	408
but the	↑	309	119	406	403
if the	↑	231	122	520	399
will be	↑	177	126	661	384
have been	↓	120	130	949	378
about the	↑	297	131	418	377
for example	↑	158	132	713	375
should be	↑	227	141	531	366
is not	↓	106	147	1042	359
there is	↓	101	514	1071	163
the same	↓	129	166	875	333
in this	↓	130	178	872	318
this is	↓	134	177	841	319

In contrast to the adjusted list from the demographic texts (Table 11), the adjusted frequency list for academic writing has both a fewer number of chunks and lower ranked chunks with central clause functions. Instead we see more of the grammatical/discourse function chunks, such as *a number of*, *in terms of*, *as well as*, *for example*. In this case the procedure, at least using these values for N_{max} and the frequency thresholds for *n*, has served to highlight the grammatical function items at the top of the list. The advantage is that it illustrates the frequent grammatical chunks (*of the*, *in the*, *with a*, *such as*, etc.) that one might tend to pass over when glancing over an n-gram list.

7 Final considerations and further developments

This paper is an initial attempt to address the recognized limitation in standard n-gram analysis when a range of values of *n* are combined, and particularly when single words and larger n-grams are combined into a single list. Because each size unit is counted on its own terms, the frequency for single words and lower values of *n* will always be larger than (or perhaps equal to) the frequency of larger units. This was clearly illustrated in the analysis of spoken language from the BNC Baby corpus with the case of the word *know*.

The adjusted frequency list procedure is presented as one possible remedy for this problem. It gives priority to larger chunks (e.g. *on the other hand*) as it builds a frequency list by not counting the included components (*on the other*, *the other hand*, *on the*, *the other*, *other hand*, *on*, *the*, *other*, *hand*). Different results are achieved by varying the size of the largest n-gram (N_{max}) at which the procedure begins and by applying different frequency (or potentially statistical) thresholds for the inclusion of specific n-grams. Three potential algorithms are presented here, two of which are used for implementation. There are likely to be other approaches as well. Future work is needed to apply the procedure to a range of corpora and to determine some criteria for determining appropriate thresholds.

The simple method presented here, along with other more complex techniques that have been recently proposed (Gries and Mukherjee 2010), demonstrates how corpus analysis continues to validate the importance of chunking in the investigation and description of language.

Notes

1. See Gries (2008) for a state-of-the-art overview of phraseological concepts in linguistic theory and computational method. The review addresses issues of terminological variation and the lack of generally accepted criteria for the identification and description of phraseologisms. Gries lays out six parameters/criteria that are designed to cover all of the aspects relevant to the notion of phraseology and by which any fully developed phraseological approach can be evaluated. He also makes suggestions with regards to methods and tools for the computational analysis of phraseology.
2. Sentence boundary punctuation has been observed in generating n-grams to produce the frequency list in Table 2. So for *...said the Mama bear. "Someone's been…* the 3-gram *mama bear someone's* would not be counted. Not all software that produces n-gram lists respects sentence boundaries in this way.
3. Most often a number of iterations of the procedure will be required to capture the appropriate size for N_{max} and the thresholds for a particular text or corpus. The procedure and suggested algorithms (see Section 4) are designed for such iterations. There are also more complex algorithms, such as the Lexical Gravity approach (most recently applied in Gries and

Mukherjee [2010]), that are designed to induce the maximum value of *n* for specific n-grams from the data.

References

Ellis, Nick C. 1996. Sequencing in SLA. Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18: 91–126.

Ellis, Nick C. 2003. Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty and M.H. Long (eds.). *Handbook of second language acquisition*, 33–68. Oxford: Blackwell.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20 (1): 29–62.

Gries, Stefan Th. 2008. Phraseology and linguistic theory: A brief survey. In S. Granger and F. Meunier (eds.). *Phraseology: An interdisciplinary perspective*, 3–25. Amsterdam and Philadelphia: John Benjamins.

Gries, Stefan Th. and Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15 (4): 520–548.

Meunier, Fanny and Sylviane Granger (eds.). 2008. *Phraseology in foreign Language learning and teaching*. Amsterdam: John Benjamins.

Nattinger, James R. and Jeanette S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

O'Donnell, Matthew Brook. 2008. KWICgrouper. Designing a tool for corpus-driven concordance analysis. In M. Scott, P. Pérez-Paredes and P. Sánchez-Hernández (eds.). *Software-aided analysis of language*. Special issue of *International Journal of English Studies* 8 (1): 107–122.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2006. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Pawley, Andrew and Frances. H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (eds.). *Language and communication*, 191–226. London: Longman.

Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, John and Anna Mauranen. 2006. *Linear unit grammar: Integrating speech and writing*. Amsterdam: John Benjamins.

Stubbs, Michael. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7 (2): 215–244.

Wray, Alison. 2008. *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Zimmer, Benjamin. 2010. On language: Chunking. *New York Times Sunday Magazine* September 19, 2010: 30.

Appendix

Text 1: Goldilocks and the Three Bears
(downloaded from http://www.dltk-teach.com/rhymes/goldilocks_story.htm)

Once upon a time, there was a little girl named Goldilocks. She went for a walk in the forest. Pretty soon, she came upon a house. She knocked and, when no one answered, she walked right in.

At the table in the kitchen, there were three bowls of porridge. Goldilocks was hungry. She tasted the porridge from the first bowl.

“This porridge is too hot!” she exclaimed.

So, she tasted the porridge from the second bowl.

“This porridge is too cold,” she said

So, she tasted the last bowl of porridge.

“Ahhh, this porridge is just right,” she said happily and she ate it all up.

After she’d eaten the three bears' breakfasts she decided she was feeling a little tired. So, she walked into the living room where she saw three chairs. Goldilocks sat in the first chair to rest her feet.

“This chair is too big!” she exclaimed.

So she sat in the second chair.

“This chair is too big, too!” she whined.

So she tried the last and smallest chair.

“Ahhh, this chair is just right,” she sighed. But just as she settled down into the chair to rest, it broke into pieces!

Goldilocks was very tired by this time, so she went upstairs to the bedroom. She lay down in the first bed, but it was too hard. Then she lay in the second bed, but it was too soft. Then she lay down in the third bed and it was just right. Goldilocks fell asleep.

As she was sleeping, the three bears came home.

“Someone’s been eating my porridge,” growled the Papa bear.

“Someone’s been eating my porridge,” said the Mama bear.

“Someone’s been eating my porridge and they ate it all up!” cried the Baby bear.

“Someone’s been sitting in my chair,” growled the Papa bear.

“Someone’s been sitting in my chair,” said the Mama bear.

“Someone’s been sitting in my chair and they've broken it all to pieces,” cried the Baby bear.

They decided to look around some more and when they got upstairs to the bedroom, Papa bear growled, “Someone’s been sleeping in my bed,”

“Someone’s been sleeping in my bed, too” said the Mama bear

“Someone’s been sleeping in my bed and she’s still there!” exclaimed Baby bear.

Just then, Goldilocks woke up and saw the three bears. She screamed, “Help!” And she jumped up and ran out of the room. Goldilocks ran down the stairs, opened the door, and ran away into the forest. And she never returned to the home of the three bears.