Exploring text-initial words, clusters and concgrams in a newspaper corpus

MATTHEW BROOK O'DONNELL, MIKE SCOTT, MICHAELA MAHLBERG and MICHAEL HOEY

Abstract

The notion of 'textual colligation' predicts that certain lexical items have a tendency to occur at particular points in a text, i.e. the beginning or end of texts, paragraphs or sentences. This paper describes new corpus-based methods developed to identify the profile of words, clusters (n-grams) and concgrams (non-contiguous patterns in variant order) in terms of their most common textual locations. Groups of co-occurring text-initial items are then analyzed in terms of their discourse function in relation to theories of newspaper structure. This analysis illustrates how methods from corpus linguistics, when targeted to specific textual positions, can complement text-linguistic analyses.

Keywords: *textual colligation, keyness analysis, concgrams, WordSmith Tools, text structure, newspaper discourse*

1. Introduction

One of the fundamental achievements of corpus linguistic method has been to demonstrate how widely the phenomenon of collocation operates across language. Within their immediate contexts, words carry with them strong preferences about which words they will occur with (Sinclair 1991, 2004; Stubbs 1996; Evert 2005). The association of particular words and phrases with certain genres or registers has been widely demonstrated (Biber 1988; Biber et al. 2007). A word may occur with similar relative frequency across two registers, say newspapers and academic writing, but have very different collocates in each. Further, different types of association (such as collocation, colligation, semantic prosody) interact or 'nest' in complex ways. The theory of Lexical Priming (Hoey 2004, 2005, 2010) introduces textual position as yet another associative feature. Hoey (2005) labels this phenomenon 'textual colligation',

Corpus Linguistics and Linguistic Theory 8–1 (2012), 73–101 1613-7027/12/0008–0073
DOI 10.1515/cllt-2012-0004

suggesting that "every word is primed to occur in, or avoid, certain positions within the discourse." (p. 13) Recent corpus studies have provided evidence that certain lexical items have a tendency to occur at particular points in a text, i.e., the beginning or end of texts, paragraphs or sentences (Scott and Tribble 2006; Hoey and O'Donnell 2007b, 2008a, 2008b; Mahlberg 2009; Csomay and Cortes 2010; Römer 2010; O'Donnell and Römer In preparation). Lexical items cannot, therefore, be assumed to be evenly distributed across texts and corpus sampling and statistical procedures should reflect this fact. Sinclair (1991: 19) long advocated 'whole text' sampling in corpus compilation in part to allow for this uneven distribution (cf. Stubbs 1996: 32).

The investigation of textual colligation necessitates a corpus made up of a large number of homogenous (in terms of text-type) whole texts.[1] The level of granularity at which one wishes to explore the association of lexis and textual position, that is, the number and size of the text units (e.g. sentence, paragraph, text), will determine the method adopted to study positional patterns. In his initial investigations of the phenomenon, Hoey began with a sentence in a specific text and then examined concordance lines from a larger corpus to ascertain how often each word or phrase occurred in sentence-, paragraph- and text-initial positions. While sufficient to substantiate the plausibility of the notion of textual colligation, clearly a more efficient and systematic approach was required. Scott and Tribble (2006: 43–52) offer some steps in that direction, still using concordances but with a more systematic statement of the textual position of each occurrence of a keyword (following Hoey's lead, they investigate *ago* in more detail).

The research reported here is one of the outcomes of a two-year project in which we aimed to test the claims for textual colligation in a rigorous and systematic manner against a large corpus.[2] We (re)applied the 'key word' procedure (Scott 1997, 2002) to compare a sub-corpus of text-initial sentences with one containing sentences that do not begin a text or paragraph. The outcomes of the comparison are lists of items that are over- and underused in those positions within newspaper texts. This procedure can be referred to as 'intra-textual keyword analysis'. These positional associations are found for single words and also for clusters.[3] For example, *yesterday* and *announced* are text-initial words in newspaper stories as are the clusters *announced yesterday*, *yesterday announced* and *it was announced yesterday*. This paper examines whether and to what extent such text-initial associations extend to non-contiguous patterns or units such as those identified by the recently proposed Concgram procedure (Cheng et al. 2006; Cheng et al. 2009). The notion of 'concgrams' allows for variation in the ordering and span of sets of co-occurring items (e.g. *under plans announced yesterday* and *plans were announced yesterday*). Using a corpus of 113,000 texts (52 million words) from the Guardian newspaper that has been divided so that all the first sentences of texts are grouped together, as are

the first sentences of paragraphs and all non text- or paragraph-initial sentences, we calculate the concgrams for each of these sub-corpora. We make use of the new WSConcGram tool in WordSmith Tools 5.0 (Scott 2008).

The primary aim of this article is to present a methodology for identifying textual colligations. Additionally, we want to make some suggestions as to how the methodology can aid in the analysis of phenomena at the interface of corpus and text analysis. We have found patterns of co-occurrence of key words within text-initial sentences. For example, *fresh,* as a text-initial key word collocates with other text-initial key words, including *controversy*, *blow*, *embarrassment*, *face|facing*, *over*, *after* and *yesterday* (and no longer with *water, fruit, food* etc.) The interpretation of such patterns suggests that the position of the key words is closely linked to their functions in texts. We have drawn on White's (1997) concept of the 'nucleus' as the textual anchor point of newspaper articles to explain such patterns (Mahlberg and O'Donnell 2008). In this article, we will return to the example of *fresh* and the nucleus pattern to show how the generation of key concgrams opens a new route to the identification of sets of words that are potential candidates for the identification of nucleus patterns.

The methodology is presented in the following steps. Section 2.1 begins with an outline of how we divide up the Guardian corpus into positional sentence sub-corpora. Section 2.2 shows the application of the intra-textual key-item analysis to discover words and clusters that exhibit associations with particular sentences in text. The results of Section 2.2 are used in 2.3 to look at whole texts and classify all the words in the text according to their text-positional associations. Section 2.4 emphasizes that textual position is not only a structural feature but is at the same time associated with functions of items in text. In particular, the finding that sets of text-initial items frequently co-occur in text-initial sentences and together play a core role in text structure and the presentation of information provides motivation for the following section. First, Section 3.1 explains the notion of a 'concgram' before Section 3.2 describes the new implementation of the WSConcGram tool. Finally, Section 4 presents some initial results of using this tool to extract text-initial concgrams.

2. A new method for discovering textual associations

The definition of the category 'text-initial' will clearly differ according to the type of text under investigation. The first chapter of a work by Dickens, for instance, could very well be counted as belonging to the beginning of the text, given the length of the complete work and the function of the opening chapter for establishing characters, settings and plot. On the other hand, in a text message of 140 characters or less, the beginning may be just the first token of the

message. This paper focuses on the associations between lexical items and the first sentences of orthographic paragraphs in a corpus of newspaper articles. The corpus consists of over 52 million words from the 'Home News' section of the Guardian newspaper from 1998 to 2004. The first step in our method is to classify sentences according to their position in text.

2.1 *Sentence classification*

The categories and labels used for the classification of each of the sentences in the articles in our corpus are as follows: 1. headline and subheadline sentences (HISC), 2. text-initial sentences (TISC) – the first non-headline sentence, 3. paragraph-initial sentences (PISC) – sentences that begin any subsequent (non-TISC) paragraph that is made up of at least two sentences, 4. single-sentence paragraph sentences (SISC) – includes both actual paragraphs of one sentence and those where white space has been used in lists, links or bullet points and 5. non-initial sentences (NISC) – any sentence in a paragraph that is not the initial sentence of that paragraph.

Table 1 below shows a newspaper text from *The Guardian* where the first part of each sentence of the text is shown along with its positional classification (in the Sub-corpus column). Sentence (1) of a text will always belong to the TISC (text-initial sentence) classification regardless of the number of sentences in this paragraph and any text preceding this first sentence is classified as HISC. Sentence (2), 'The proposal, which was put into the bill without consulting . . .', begins the second paragraph (P2), which has 2 sentences in it, so receives the PISC classification. Sentence (7), 'The clause says the home secretary may by order limit the release of . . .', occurs in the fourth paragraph of the text (P4) and is the second sentence of its paragraph, so it is classified as NISC (a non-initial sentence). While sentence (12), 'The proposals have angered the Liberal Democrats and Robert . . .', is also the first sentence of a paragraph (P7) it is the only sentence in the paragraph, so it receives a SISC classification.

The number of sentences in the TISC (Text-Initial Sentence Corpus) category (113,288) is equivalent to the number of articles in the whole corpus, as every article will have at least one paragraph and one sentence. While a discussion of the issues surrounding corpus size is outside the scope of this paper, we would highlight two points. First, a corpus of over 52 million words of a specific (sub-)register can be viewed as of satisfactory size for generalization and the identification of lexico-grammatical patterns. Second, a corpus made up of over one hundred thousand complete texts offers considerable opportunities for the examination of text-linguistic features – and specifically lexical patterns in text-initial sentences – from a quantitative perspective.

Table 1. *Classification of sentences in a newspaper text according to textual position*

Paragraph	Sentence		Sub-corpus
		Backlash over Straw veto on release of information[4]	*HISC*
P1	(1)	The home secretary, Jack Straw, waded into a fresh bout of controversy over his much-criticised freedom of information bill yesterday when it emerged he had given himself a new veto to ban the release of any public documents.	*TISC*
P2	(2)	The proposal, which was put into the bill without consulting . . .	*PISC*
	(3)	The bill was already certain to face a rough ride through parliament . . .	*NISC*
P3	(4)	Mark Fisher, the former arts minister sacked by Tony Blair in his . . .	*PISC*
	(5)	It basically allows Jack Straw to put a razor through the whole bill . . .	*NISC*
P4	(6)	The loophole was only spotted by campaigners and MPs last week . . .	*PISC*
	(7)	The clause says the home secretary may by order limit the release of . . .	*NISC*
P5	(8)	Maurice Frankel, director of the Campaign for Freedom of . . .	*PISC*
	(9)	It is also giving the same power to any successor government.'	*NISC*
P6	(10)	MPs are particularly angry that Mr Straw is proposing to use an . . .	*PISC*
	(11)	This will allow him to ban release of any particular piece of . . .	*NISC*
P7	(12)	The proposals have angered the Liberal Democrats and Robert . . .	*SISC*
P8	(13)	Mr Maclennan said last night: 'This amounts to a dispensing clause.	*PISC*
	(14)	It basically says we shall apply the law, except when Jack Straw . . .	*NISC*
	(15)	This seems to be based on excessive caution.'	*NISC*
P9	(16)	Ann Widdecombe, the shadow home secretary, has also sought to . . .	*SISC*
P10	(17)	The home office claimed the new clause was necessary to modernise . . .	*PISC*
	(18)	A spokeswoman said: 'If a future home secretary used this clause to . . .	*NISC*

Table 2. *Positional sub-corpora from the Guardian Home News corpus (1998–2004)*

	Positional Sub-corpora			
	TISC	PISC	SISC	NISC
tokens	3,122,037	12,521,902	17,129,694	19,338,590
types	58,432	127,038	137,322	141,793
sentences	113,288	607,125	555,641	1,064,493
mean (in words)	28	21	31	18

2.2 *Intra-textual keyness*

The positional categories (TISC, PISC, etc.) can either be treated as a type of annotation added to sentence (`<s>`) elements in each XML document (e.g. `<p num="1"><s num="1" posCat="TISC">...</s>`) or be used to create sub-corpora where each file containing an article is split into 5 separate files corresponding to the 5 positional categories. It should be clear that these are merely notional and methodological differences. We have used both strategies in the course of the project, but found that the second one (separate sub-corpora) was more practical for use with WordSmith Tools (Scott 2008). Word and cluster (or n-gram) frequency lists were created from the sentences of each of these categories and pair-wise comparisons are made between the resulting lists using the KeyWords procedure (Scott 1997).[5] The outcome of this procedure is a matrix that classifies each item (word or group of words) found to be 'key' in at least one of the comparisons, indicating whether it has a significant tendency to occur in various locations in text, namely, Text-Initial, Paragraph-Initial and Non-Initial positions. Figure 1 illustrates the six comparisons between TISC, PISC and NISC used to construct this matrix.

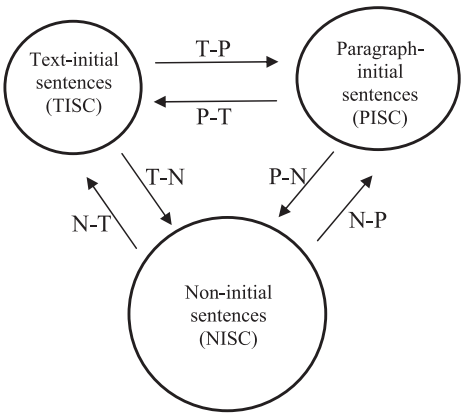

Figure 1. *Intra-textual keyness comparisons between sub-corpora*

The circles represent the frequency lists of words and clusters (n-grams) generated from each of these sub-corpora. The arrows indicate the direction of the keyness comparison. For instance, T-N represents the comparison between items from TISC against those in NISC. Items highlighted as positively key from the T-N comparison can be said to have an association with text-initial sentences, based upon the statistically significant relative frequency of the item in TISC compared to its relative frequency in NISC.

2.2.1 *Textual position matrix*

To construct the textual-position matrix, the union of all resulting key item lists is taken and then each item is checked against each of the lists. If the item turns up as key on a list, a 'Y' is placed in the appropriate column and an 'N' if it is not on the list. A look at Tables 3 and 4 should help to clarify the preceding description. Table 3 shows 15 individual words from the 4,467 words found to be key in at least one of the comparisons illustrated in Figure 1. The examples in Table 3 do not appear in any particular order. The first word in the table, *force*, has the pattern YNNNNN, which means it is key only in the TISC against NISC keyword comparison and is not found on any of the other lists. This pattern can be interpreted as a weak text-initial sentence preference. Item 5, *spy*, has the pattern YYNNNN as it occurs as positively key in both the TISC against NISC and TISC against PISC keyword comparisons and on none of the other lists. This pattern is interpreted as indicating a strong

Table 3. *Fifteen sample words and their text positional association patterns*

#	Item	Key Word comparison lists					
		T-N	T-P	P-N	P-T	N-T	N-P
1	FORCE	Y	N	N	N	N	N
2	LATER	N	N	N	Y	Y	Y
3	ENVIRONMENT	N	N	Y	N	N	N
4	INCIDENTS	N	N	N	Y	N	N
5	SPY	Y	Y	N	N	N	N
6	FRESH	Y	Y	Y	N	N	N
7	GREETED	N	N	Y	N	N	N
8	HARDLINE	Y	N	Y	N	N	N
9	MOVE	N	N	Y	Y	N	N
10	DISAPPEAR	N	N	N	N	N	Y
11	PROUD	N	N	N	Y	Y	Y
12	NATURE	N	N	N	Y	Y	N
13	DESERVE	N	N	N	N	Y	Y
14	SPOKESMAN	N	N	Y	Y	Y	N
15	THINKTANK	Y	Y	Y	N	N	N

preference for occurring in text-initial sentences (in Guardian Home News articles).

The six-way comparison between the three positional sub-corpora allows distinctions such as strong versus weak text-initial preference to be made. If an item is key in T-N but not T-P it means that relatively it is more frequent in text-initial sentences than in non-initial sentences but not so in comparison to the first sentence of non-initial paragraphs (PISC). A pattern of YNYNNN, that is key in T-N and P-N, such as that for item 8, *hardline*, would indicate a strong paragraph-initial (both text-initial and other paragraphs) tendency without a preference for either a text-initial or non text-initial paragraph.[6] The word *fresh*, which we discuss in more detail in Section 2.4, exhibits a YYYNNN pattern. That is, *fresh* is key in T-N, T-P and P-N comparisons, demonstrating the strong text-initial association of *fresh* in the Guardian corpus along with a weaker preference for other paragraph-initial positions over non paragraph-initial sentences. Item 11 in Table 3, *proud*, exhibits the pattern NNNYYY, as it is not key in the Text-initial Sentence Corpus (TISC) against the Non-initial Sentence Corpus (NISC) (T-N), TISC against the Paragraph Initial Sentence Corpus (PISC) (T-P) or PISC against NISC (P-N) comparisons but is key for comparisons of PISC against TISC (P-T), NISC against TISC (N-T) and NISC against PISC (N-P). How is such a pattern interpreted? First, it is clear that *proud* is not a text-initial word in our Guardian Home News corpus (from the 'N' values for T-N and T-P and the 'Y' values for P-T and N-T). Normalized frequencies confirm these findings in that occurrences of *proud* per million words are TISC: 16.0, PISC: 32.3 and NISC: 49.9. Secondly, it is not a paragraph-initial word, although when it is paragraph-initial it is more likely to occur in a non text-initial paragraph.

A similar matrix was constructed for the union of 2 to 5 word clusters found to be key in at least one of the comparisons illustrated in Figure 1. The cluster matrix contains 50,861 items and Table 4 shows again 15 examples to illustrate the output of the method and the interpretation of the patterns.

Clusters with strong text-initial associations (YYNNNN pattern) include items 1 (*Israeli Prime Minister*), 2 (*a man whose body was*), 4 (*in a move likely*), 6 (*the government was last night*), 8 (*at their home*) and 14 *(face a)*. Strongly non text-initial items include the trigram *is not known* and the bigram *or six* (items 12 with NNNYYN patterns). Two items, 3 (*move follows*[7]) and 9 (*the figures are*), have a NNNYNN pattern where the only significant positional association is from the P-T comparison, suggesting that they are more likely to be found in non text-initial paragraphs when they are in the first sentence of a paragraph. The trigram *the figures are* does not occur at all in TISC, that is, in our Guardian Home News corpus *the figures are* is never found as part of the first sentence of any of the 113,288 articles. In contrast it occurs 71 times (5.6 per million words) in PISC and 67 times (3.6 per million words) in

Table 4. *Fifteen sample clusters (2–5 words) and their text positional association patterns*

#	Item	Key Cluster comparison lists					
		T-N	T-P	P-N	P-T	N-T	N-P
1	ISRAELI PRIME MINISTER	Y	Y	N	N	N	N
2	A MAN WHOSE BODY WAS	Y	Y	N	N	N	N
3	MOVE FOLLOWS	N	N	N	Y	N	N
4	IN A MOVE LIKELY TO	Y	Y	N	N	N	N
5	AND TOOK	N	N	N	N	Y	N
6	THE GOVERNMENT WAS LAST NIGHT	Y	Y	N	N	N	N
7	A CAMBRIDGE UNIVERSITY	N	Y	N	N	N	N
8	AT THEIR HOME	Y	Y	N	N	N	N
9	THE FIGURES ARE	N	N	N	Y	N	N
10	FOR A PINT	N	N	N	N	N	Y
11	LAST NIGHT DENIED THAT	N	N	Y	N	N	N
12	IS NOT KNOWN	N	N	N	Y	Y	N
13	WILL BE ABLE TO	N	N	N	N	N	Y
14	FACE A	Y	Y	N	N	N	N
15	OR SIX	N	N	N	Y	Y	N

NISC. Item 11, the four-gram *last night denied that,* has a NNYNNN pattern or weak paragraph-initial preference association. It is a relatively low frequency item with 3 hits in each of TISC and NISC and 14 in PISC (frequencies per million words are TISC: 0.96, PISC: 1.12, NISC: 0.16).

The method outlined above and the resulting matrix relating items to their textual position associations achieve the goal of providing a systematic and comprehensive account of the associations that hold between items (in terms of 1 to 5 word clusters/n-grams) and particular sentence positions within newspaper text. Even with the high threshold for significance used in our key-item comparisons (using the log likelihood with $p < 0.000001$) our analysis revealed 4,467 individual words which were key in at least one of the comparisons between TISC, PISC and NISC illustrated in Figure 1. Patterns associated with text-initial position (YYNNNN, YNNNNN and YYNNNY), account for 1,600 (36%) of the total key words and 29,303 (58%) of the key clusters. The patterns associated with paragraph-initial position (NNYYNN, NNYNNN and NNNYNN) account for a further 732 (16%) of the key words and 5,755 (11%) of the key clusters. The patterns associated with non-initial position (NNNNYY, NNNYYY and NYNNYY) make up only 486 (11%) of the key words and 3,105 (6%) of our key clusters. Other patterns are associated with relatively few words and clusters. In our corpus, 4,467 key words is roughly one out of every forty types. This suggests that textual colligation is not a phenomenon limited to a few peculiar words such as *ago* (which has been thoroughly explored by Hoey [2004, 2005]; also Scott and Tribble 2006) nor to well known text-initial

clusters such as *once upon a time*. The findings provide support to the claim made by Lexical Priming theory that lexical associations extend to the level of text.

2.3 *Using text positional matrix data to examine individual texts*

The method discussed in the previous section, which necessitates the use of a corpus made up of full and not partial texts, illustrates the confluence of corpus linguistic and text linguistic methods and approaches (cf. Partington 2004; Baker 2006). In such approaches there is a movement back and forth between the close and detailed analysis of individual texts and the observation of localized (lexical) patterns and associations from a large corpus of such texts. The textual position matrix is an example of the latter, but the items it contains can be used to examine an individual text by marking each of the words according to their text-positional associations. In the interests of space, below we show just the first three sentences coded in this way from the newspaper text shown in Table 1. Words marked in bold exhibit **text-initial sentence** associations, those in italics *paragraph-initial* (non text-initial paragraphs) and underlined words have associations with non paragraph-initial sentences.

(1) *The* **home secretary**, **Jack Straw**, waded **into a fresh** bout **of controversy over** his much criticised freedom **of** information ***bill*** **yesterday when** it **emerged** he had given himself **a new** veto to **ban** *the* release **of** any **public** ***documents*** (TISC).

(2) *The proposal*, *which* was put into *the* ***bill*** without consulting Parliament, united *in opposition* Labour rebel *MPs*, Liberal Democrats and Conservatives on the eve of *the* measures facing scrutiny *in the* Commons (PISC).

(3) *The* ***bill*** was already certain to **face a** rough ride through Parliament because *Mr* Straw had blocked reforms to allow the new information commissioner power to override **ministers** to order ***publication*** of ***contentious*** information and *insisted* on wide-ranging exclusions to *the* release of documents, particularly *in* policy-making areas (NISC).

Of the 38 words in the first sentence (1) of the text (a TISC sentence) 21 (55%) exhibit a text-initial sentence association compared with 3 that have a paragraph-initial association and just 2 with a non-initial sentence association. In contrast, sentence (2) (a PISC sentence), contains only one word, *bill*, that was found to have a text-initial sentence association. It is highlighted in both bold and italic because it also exhibits a general paragraph-initial preference. Of the 32 words in this sentence, 10 (31%) exhibit a paragraph-initial association. The third sentence (3) (a NISC sentence) contains 48 words, 6 of which

have text-initial associations, 8 with paragraph-initial and only 3 that show a pattern in our matrix that we interpret as a non-initial association. This may at first seem to raise questions as to the value of the intra-textual key-item analysis and the data in our matrix. However, from Table 1, we can see that NISC is made up of roughly ten times as many sentences as TISC and nearly twice as many as in PISC. Further, in terms of text structure non-initial sentence position is the least prominent or distinct. So it seems reasonable that the strongest associations will form between the less frequent (for TISC just one per article) and more prominent sentences of a newspaper article.

If the vocabulary (types) in our Guardian Home News corpus were distributed at random and evenly across all the sentences of all of the articles, we would not expect to see such a pattern. The Guardian writers may have acquired these associations over time and when they (and their editors) come to write the first sentence of a Home News piece they can be said to be 'primed' to select words that are statistically over-represented in TISC. Equally, readers of the Guardian may, over time, acquire these word-to-text-position associations in the same way that they acquire collocations, semantic associations and semantic prosodies.[8]

We now move on from the text-positional associations of individual words to associations between strings of 2, 3, 4 and 5 words and textual position. Below the first three sentences of the above text are displayed again using the same coding (bold = text-initial, italic = paragraph-initial and underlined = non-initial) with only items of 2 words or more (up to 5 grams) highlighted.

(1) **The-home-secretary-Jack-Straw**, waded **into-a-fresh** bout of **controversy-over** his much criticised **freedom-of-information-bill yesterday-when-it-emerged** he-had given himself **a-new** veto **to-ban** the release of any public documents (TISC).

(2) *The-proposal*, *which-was* put into *the-bill* without consulting Parliament, united in opposition Labour rebel MPs, Liberal Democrats and Conservatives on **the-eve-of** the measures facing scrutiny in the Commons (PISC).

(3) *The-bill* was already certain **to-face-a** rough ride through Parliament because-*Mr*-*Straw* had blocked reforms to allow *the-new* information commissioner power to override ministers to order publication of contentious information and insisted on wide-ranging exclusions *to-the* release of documents, particularly in policy-making areas (NISC).

The first sentence contains 7 clusters (*The-home-secretary-Jack-Straw*, etc.) that exhibit a pattern in our intra-textual key-item analysis that we interpret as a text-initial association, compared with only one such cluster in each of the following two sentences. The use of clusters or n-grams as the unit of analysis

Table 5. *Elements of the nucleus pattern*

A	B	C	D	E	F	G	H	I
X	TIME	FACE SUFFER *sparked* *triggered* *embroiled in* etc.	***fresh***	*controversy* *row* *embarrassment* *blow*	*over* *about*	**Y**	*when* *after*	**Z**

appears to reduce some of the 'noise' inherent in the use of isolated words and reinforces Hoey's notion of 'nesting', that is, word-to-word associations (collocations) in turn form further associations, be they semantic, syntactic, pragmatic or textual.[9]

2.4 *Looking at text- and paragraph-initial items in context*

The use of the text-positional matrix to annotate individual words in a text is in the first instance a descriptive exercise. What is now needed is an explanation for why these items seem to occur together. This question addresses the local functions of the words in texts.[10] Mahlberg and O'Donnell (2008) start with the textual-initial word *fresh* and from a KWIC analysis build up a description of a nucleus pattern, after White (1997) who describes newspaper texts in terms of 'nucleus' (first sentence + headline material) and a series of following 'specifications' (see Appendix). Such a nucleus pattern has introductory functions highlighting the newsworthiness of a news item at the beginning of the article. The text-initial *fresh* pattern can be represented as shown in Table 5 (**X**, **Y** and **Z** relate to participants and events).

The following examples are TISC sentences that are instances of this pattern (Example 4 is the first sentence of the example article used above and shown in Table 1):

(4a) [The home secretary, Jack Straw,]^{A} [waded into]^{C} a [fresh]^{D} [bout of controversy]^{E} [over]^{F} [his much-criticised freedom of information bill]^{G} [yesterday]^{B} [when]^{H} [it emerged he had given himself a new veto to ban the release of any public documents]^{I}.

(5a) [The government]^{A} was [yesterday]^{B} [embroiled in]^{C} a [fresh]^{D} [row]^{E} [over]^{F} ['fat cat' pay rises]^{G} [after]^{H} [it emerged that Chris Woodhead, Chief Inspector of Schools, is to be reappointed for a further five-year term on a significantly higher salary.]^{I}

(6a) [Labour]^{A} [faced]^{C} [fresh]^{D} [embarrassment]^{E} [over]^{F} [its funding]^{G} [yesterday]^{B} [when]^{H} [the independent electoral commission criticised it for continuing to break rules on donations.]^{I}

(7a) [The government]^{A} was [yesterday]^{B} [at the centre of]^{C} a [fresh]^{D} [row]^{E} [over]^{F} [plans for a shake-up in the way history is taught in schools]^{G}, [after]^{H} [the Tories joined forces with traditionalist history experts to accuse Labour of undermining Britain's national identity.]^{I}

The textual function of this pattern is explored in detail in Mahlberg and O'Donnell (2008). For example, they found that 'controversy nouns' (*controversy*, *row*, *embarrassment*, *blow*) carry evaluative meaning and highlight newsworthiness while at the same time they are general enough to cover a variety of subject matters (Francis 1994 calls such nouns 'Labels'). The above examples only display the first sentence of each article, but it is important to emphasize that the functions of the text-initial words cannot be described with regard to this one sentence alone. The nucleus encapsulates the main meanings that will be unfolded in the article. It serves as the textual 'anchor point'. Thus subsequent sentences relate back to the nucleus and are "organized 'orbitally' rather than linearly" (White 1997: 116; see analysis of example text in the Appendix).

Mahlberg and O'Donnell (2008) is based on a small set of articles that are hand-annotated according to the components of the nucleus pattern. However, with the help of the text-positional matrix data, the nucleus pattern can be described in further detail. Below are the examples of the nucleus pattern with text-initial sentence association highlighted in bold.

(4b) [The **home secretary**, **Jack Straw**,]^{A} [waded **into**]^{C} **a** [**fresh**]^{D} [bout **of controversy**]^{E} [**over**]^{F} [his much-criticised freedom **of** information **bill**]^{G} [**yesterday**]^{B} [**when**]^{H} [it **emerged** he had given himself **a new** veto to **ban** the release **of** any **public documents**]^{I}.

(5b) [The **government**]^{A} was [**yesterday**]^{B} [**embroiled in**]^{C} **a** [**fresh**]^{D} [**row**]^{E} [**over**]^{F} ['fat **cat**' **pay rises**]^{G} [**after**]^{H} [it **emerged** that Chris Woodhead, Chief Inspector **of Schools**, is to be reappointed **for a** further five-year term **on a** significantly higher salary.]^{I}

(6b) [**Labour**]^{A} [**faced**]^{C} [**fresh**]^{D} [**embarrassment**]^{E} [**over**]^{F} [**its** funding]^{G} [**yesterday**]^{B} [**when**]^{H} [the **independent** electoral commission criticised it **for** continuing to **break** rules **on** donations.]^{I}

(7b) [The **government**]^{A} was [**yesterday**]^{B} [**at the centre of**]^{C} **a** [**fresh**]^{D} [**row**]^{E} [**over**]^{F} [**plans for a** shake-up **in** the way **history** is taught **in schools**]^{G}, [**after**]^{H} [the Tories joined **forces** with traditionalist **history** experts to accuse **Labour of** undermining Britain's **national** identity.]^{I}

The large number of words with strong text-initial associations in these sentences is clear from the highlighting. While the nucleus pattern as such is of the type that relies on detailed manual observations for its discovery, once

established it can help the functional interpretation of items identified by the intra-textual keyness comparisons. One aspect of interest is the elevated nature of much of the text-initial vocabulary. That is, an issue or event is *fresh* (YYYNNN) and not just *another* (NNNYYN) or a *further* (NNNYYN) row or controversy, and individuals and governments are *plunged into* (YYNNNN) or *embroiled in* (YYNNNN) controversy rather than just becoming *involved* (NNNYYN) in such. Further, in terms of White's orbital structure analysis, many of the text-initial items in the nucleus function as the 'anchor points' that are targets for linking from the following specifications. For example, *the proposal(s)* and *measure* (both NNYYYN), *angered* (NNYNNN) and *angry that* (NNNYYN) in specifications 1, 4, 5 and 6 relate back to elements of the nucleus pattern (see Appendix).

A comprehensive capture and analysis of nucleus patterns is as yet beyond the scope of automatic extraction. At the same time, there might be a way to capture the co-occurring key word sets that make up each of the slots in a pattern (e.g. FACE + *fresh* + *controversy* + *over* + *when*) more directly. To explore this potential, the remainder of this paper makes use of the concept of a concgram and its implementation in WordSmith Tools Version 5.0.

3. Concgrams and tools for concgram extraction

3.1 *Concgrams and ConcGram*

A primary strand of corpus linguistic research from the pioneering work reported in the OSTI Report (Sinclair et al. [1970] 2004) through to the most recent studies of phraseology (Granger and Meunier, 2008; Römer and Schulze, 2009) is the identification of co-occurring items within a specified window (e.g. non-consecutive items like *strong* as a collocate of *tea*) or immediate consecutive clusters (e.g. *at the end of* and *Merry Christmas*). Virtually all corpus software can easily extract and display such patterns, but there are some limitations. Non-consecutive collocations are restricted to pairs of words and clusters are of fixed order (e.g. *yesterday he announced*, *he yesterday announced* and *he announced yesterday* would be counted as three separate n-gram types). The Concgram procedure takes a whole corpus of text and finds all sorts of combinations like these, whether consecutive or not, and is able to group the resulting patterns that share the same items but vary in their ordering. Cheng et al. (2006: 414) define a concgram as "all of the permutations of constituency variation and positional variation generated by the association of two or more words." This means that the associated words comprising a paticular concgram may be the source of a number of 'collocational patterns' (Sinclair et al. 2004: xxvii).

The Concgram procedure attempts to move away from the node-centred pair-wise investigation of association, that is, selecting a word and collecting frequently occurring collocates within a set span one at a time. There is still the need for a starting point, or what the procedure terms an 'origin'. The ConcGram program (Greaves 2009) adopts an iterative search-based approach to the collection of concgrams. Beginning with a single word as the origin, it searches for all the concgrams of length 2 within a specific window. Then each of these in turn become an origin of length 2 to search for all the concgrams of length 3 within the span. This is repeated with an origin of length 3 to find all concgrams of length 4, and so on (Greaves 2009: 15). Greaves (2009) provides an introduction to both the concept of concgrams and to using the ConcGram program. The flexibility and descriptive power of a concgram is that it captures two types of variation. Constituent variation allows for items to be at varying distances from each other and for the same item to have different syntactical functions but still be counted as an instance of co-occurrence. For example, *announced plans* and *announced the revision of plans* are instances of a single concgram of length two for the items *announced* and *plans*. The grouping of these two is not possible by collecting n-grams. The second kind of variation concerns the order of the items in the concgram. This is variation of the kind: *announced plans* and *under plans announced*. While this kind of ordering variation can be captured for pairs of words using the standard collocational window technique it is not possible for items of 3+ word length and again cannot be captured in n-grams.

The potential of extracting concgrams for the analysis of text- and paragraph-initial patterns should be clear given the description of the *fresh* nucleus pattern in the previous section and the examples of text-initial key-items throughout the paper.

3.2 *WSConcGram*

WSConcGram is an evolving implementation of the Concgram concept in WordSmith Tools 5.0 (Scott 2008). It is designed to find concgrams efficiently.[11] WSConcGram builds upon the index format available in WordSmith since Version 4.0 which is used to extract clusters/n-grams from a corpus. This index is augmented so that it is possible to find all pairs, triples, quadruplets, quintuplets, etc., within a specific span above a given threshold. Table 6 shows examples of each of these related to the previously discussed *fresh* nucleus pattern.

The first step in generating concgrams for a corpus using WSConcGram is to create an index file in the WordList tool and to specify the restrictions (e.g. whether to observe sentence boundaries) and span (see Figure 2).

These will impact the pairs, triples and so on, that can be collected in subsequent processing. WSConcGram then carries out two further stages of

Table 6. *Examples of concgrams of size 2, 3, 4 & 5 extracted from TISC*

PAIRS	*fresh controversy*, *government fresh*, *fresh over*, *controversy over*, *faced over*, *government faced*, *controversy after*, *fresh yesterday*
TRIPLES	*fresh controversy over*, *government fresh over*, *faced over after*, *faced fresh yesterday*
QUADRUPLETS	*government faced fresh controversy*, *government fresh controversy after*
QUINTUPLETS	*government faced fresh controversy after*, *fresh controversy yesterday over after*

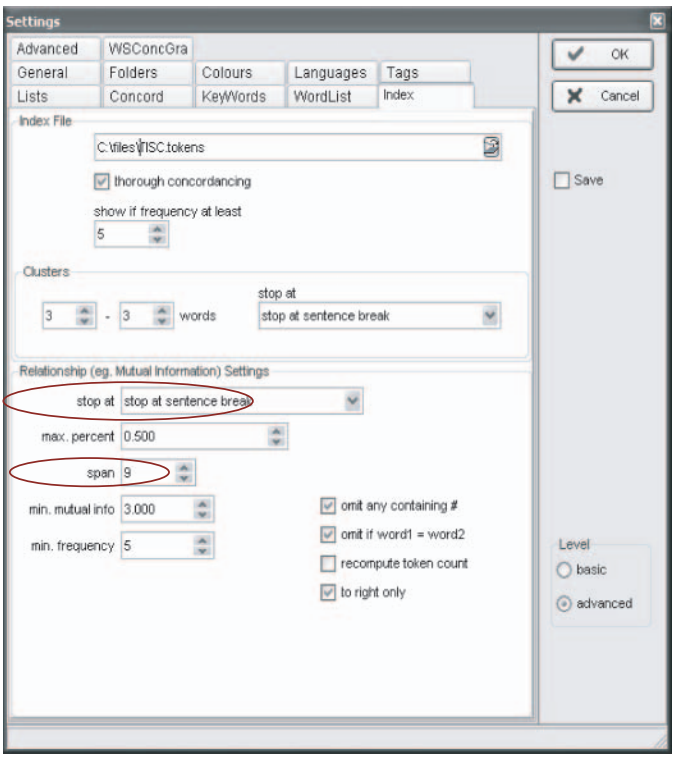

Figure 2. *Index settings in WordSmith 5.0 for generating concgrams*

processing, first to generate all the base pairs of collocating words within the set span (generating an index file with the extension `.base_pairs`). The default frequency for co-occurrence is 5 times, but this can be altered in the WSConcGram settings (Figure 3).

The key difference between the WSConcGram and the ConcGram implementation of the Concgram concept is the insight that once you have generated

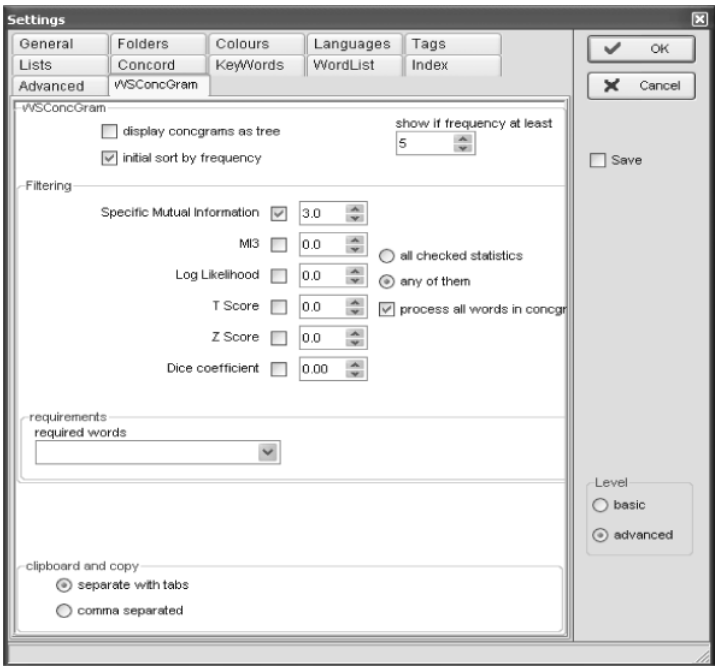

Figure 3. *Settings for WSConcGram tool in WordSmith Tools 5.0*

all co-occurring pairs (2 concgrams) it is then possible to generate all 3, 4, 5, 6, and upwards, concgrams without having to go back to the corpus and carry out further searching based on the origins (see Greaves 2009: 15). This makes it feasible to collect all the concgrams in a corpus of 3 or 4 million words in minutes. The second step of WSConcGram processing generates the final concgram index (with file extension `.base_index_cg`) that is used by WSConcGram to view and query the concgram index. The default view shows all the words involved in concgrams ordered by concgram token frequency. These are the single word origins. For instance, Figure 4 shows the most frequent words in the concgrams generated for text-initial sentences from our Guardian Home News corpus. Unsurprisingly, the most frequent type is *the* which is found in 506,269 concgram tokens, followed by *of* (292,985), *to* (280,826) and *a* (277,067).

One or more words from this left-hand list can be selected as origin items and the concgrams in which they occur are displayed in the right hand pane. The concgrams are either grouped hierarchically in a tree structure or as a flat list. Figure 5 shows part of the concgram tree for the 2081 concgrams containing *fresh* from TISC. Similar items are grouped together and each branch of the

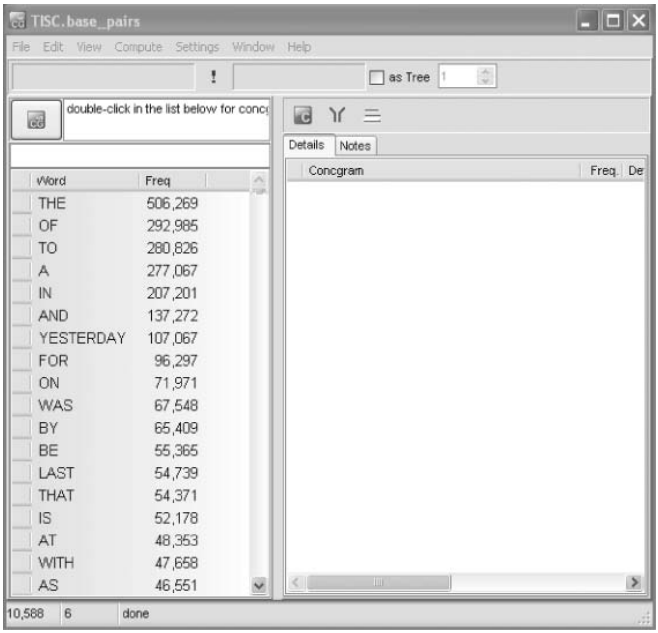

Figure 4. *Viewing concgrams in WSConcGram tool of WordSmith Tools 5.0*

tree shows how many sub-items and how many items of its own it has. For example, the node *for a* has (+3) items and sub-items: 1 occurrence of *fresh at plans for a* and 2 of *fresh for a*. The latter is itself a further node in the tree. At first it may seem confusing, but (+12) 2 indicates there are 14 tokens related to the node *fresh for a.* The procedure of building the concgram index looks for the longest (maximal) concgrams possible at each stage, which reduces the count for smaller concgrams that contain some of the same items. For example, the 5 concgram *are braced for a fresh* occurs once. The 4 concgram *braced for a fresh* has two occurrences, but neither of them is the one in 5 concgram *are braced for a fresh*. The other nine 4 and 5 concgrams, each occurring one time, that include *for a* are: *him a for fresh*, *have called for a fresh*, *Ireland is for a fresh*, *this morning for a fresh*, *issued a fresh apology for*, *be balloted for a fresh*, *could be for a fresh*, *give him a for fresh*, *fresh at plans for a*. These reduce the count for the 3 concgram *fresh for a* by 12 which means there are just two tokens of *fresh for a* as a maximal concgram.

Any concgram(s), at any level in the hierarchy, can be displayed in context through the Concord tool. In both the list and tree views concgrams are presented in context order (compare *a* in *braced for a fresh* and *give him a for fresh*) which aids in their (partial) interpretation out of context.

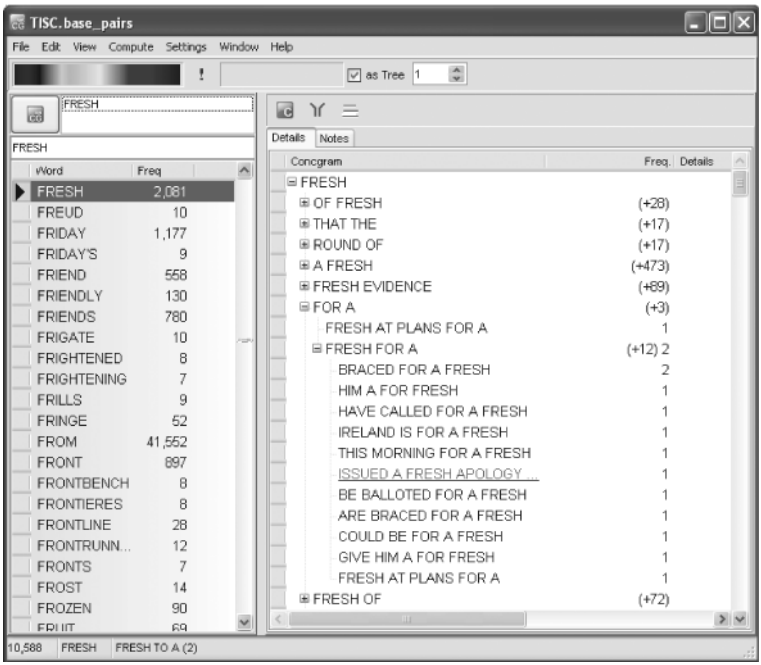

Figure 5. *Exploring concgrams for* fresh *in WSConcGram tool of WordSmith Tools 5.0*

4. *fresh* concgrams and candidates for the nucleus pattern

The final section of this paper explores text-initial concgrams and focuses particularly on those containing *fresh*. We extend the use of the KeyWords tool to compare the frequency of different concgrams in TISC and NISC. That is, we examine key concgrams, sets of co-occurring words found with a significantly higher relative frequency in text-initial sentences than in non-initial sentences.

4.1 *Procedure for extracting text-initial concgrams*

Two indexes were built for TISC and a sample of NISC (around 7.5 million words) using a span of 10 words and stopping at sentence boundaries. Using a 10 word window adds considerably to the computational task of collecting co-occurring items, but as can be seen in the examples for the *fresh* nucleus pattern (see Section 2 above and Mahlberg and O'Donnell 2008) frequently co-occurring text-initial items can be spread across the whole sentence. As the

mean length of sentences in TISC is 27.56 (compared to 18.13 for NISC) a large window is warranted.

For TISC, the WordSmith index contains 3.12 million tokens, 58,432 types and 113,288 sentences. Applying the WSConcGram procedure to this index results in 13,473 single origin items. For NISC the WordSmith index contains 7,559,300 million tokens, 94,874 types and 417,010 sentences. Applying the WSConcGram procedure to this index results in 20,226 single origin items. As mentioned in Section 3, WSConcGram collects maximal (longest) concgrams which may be too specific and low frequency for the kind of comparison between text-initial and non-initial we are interested in here. It is possible to export the concgram lists as text files. These files were processed using a Python script to extract smaller sets of co-occurring items for specific words from the TISC and NISC concgram lists that could then be compared in a key-item analysis.

4.2 *Text-initial concgrams containing fresh*

Table 7 shows the top 30 concgrams containing *fresh* of length 4 and greater from TISC (left hand column) and NISC (right hand column). The five underlined concgrams are shared in the top 30 of both concgram lists. The words marked in bold in the TISC concgram list are text-initial keywords from the analysis described in Section 2.

These lists can be turned into WordSmith word lists and used in the KeyWords tool.[12] Table 8 shows a selection of 20 text-initial (positively) key concgrams containing at least four items.

Many of these items fit the *fresh* nucleus pattern shown in Table 5 (Section 2.4); for example, *fresh after last night, fresh last night when, fresh controversy last night, fresh embarrassment last night* all illustrating elements as indicated in Table 5 (Section 2.4). Although detailed examination of each concgram in context is required to explore the specific functions of the items in text, the main point to be made here is that the use of concgrams and key-item analysis does provide a method of discovering sets of such items in a single step. While the above concgrams were extracted for the origin *fresh*, other origins also provide sets of candidates that suggest a nucleus pattern. Starting with *controversy* and *row,* two of the nouns found to collocate with fresh in text-initial sentence (see Section 2), we find the examples below. In both cases there are concgrams containing *fresh,* but also concgrams without *fresh*.

4.2.1 *Text-initial concgrams with origin* ***controversy***

The following concgrams built around the origin word *controversy* are key for their frequency in text-initial sentences:

Table 7. Fresh *concgrams of length 4+ extracted from TISC and NISC*

TISC concgrams		NISC concgrams	
fresh ***a last night***	148	*fresh a of the*	51
fresh ***last night*** *the*	122	*fresh in of the*	51
fresh ***a of*** *the*	116	*fresh and of the*	43
fresh ***after last night***	104	*fresh a the to*	39
fresh ***last night when***	77	*fresh a air of*	35
fresh ***last night*** *was*	73	*fresh air breath of*	35
fresh ***last night*** *to*	71	*fresh a in the*	29
fresh ***a*** *the to*	69	*fresh a for the*	28
fresh the ***when yesterday***	64	*fresh be to will*	26
fresh ***into last night***	64	*fresh a at the*	25
fresh ***a in*** *the*	60	*fresh of the to*	24
fresh ***last night of***	59	*fresh is of the*	24
fresh ***in last night***	58	*fresh of on the*	24
fresh ***a night*** *the*	56	*fresh and the to*	23
fresh ***a row*** *the*	55	*fresh the to will*	23
fresh ***a*** *the* ***yesterday***	53	*fresh be the will*	23
fresh ***a over row***	53	*fresh a be to*	23
fresh ***last night*** *that*	51	*fresh a and of*	23
fresh ***last night over***	49	*fresh a is of*	23
fresh ***of*** *the* ***yesterday***	46	*fresh a start the*	22
fresh ***a of*** *to*	46	*fresh air of the*	22
fresh ***controversy last night***	46	*fresh a is the*	21
fresh about ***of*** *the*	45	*fresh a from the*	21
fresh ***a last*** *the*	44	*fresh for is the*	21
fresh ***in of*** *the*	44	*fresh for start the*	21
fresh ***a night*** *to*	44	*fresh a and to*	21
fresh ***of*** *the to*	43	*fresh a new start*	21
fresh evidence ***of*** *the*	43	*fresh and from the*	20
fresh ***a when yesterday***	41	*fresh for in the*	20
fresh it ***last night***	40	*fresh in the to*	19

controversy that when yesterday (16-2), *controversy a after last night* (13-0), *controversy fresh last night over* (9-0), *controversy after emerged it night* (5-0)

(8) **Controversy** over the national lottery raged anew **yesterday when** regulators revealed **that** Camelot was expected to miss its good causes goal by up to £5bn.

Example (8) is similar to the examples on the basis of which we established the nucleus pattern. There is the evaluative noun (*controversy*), the time reference (*yesterday*), and the conjunction (*when*). The verb *rage* shares qualities with the verbs in the nucleus pattern as it indicates high impact; and the adverb

Table 8. *Text-initial key concgrams of length 4+ containing* fresh

	TISC freq.	NISC freq.	Keyness
fresh after last night	104	0	92.41
fresh last night when	77	0	68.41
fresh the when yesterday	64	0	56.86
fresh a row the	55	0	48.86
fresh a over row	53	0	47.08
fresh last night over	49	0	43.53
fresh controversy last night	46	0	40.86
fresh faced last night	39	0	34.64
fresh a blair tony	39	0	34.64
fresh over the yesterday	38	0	33.75
fresh embarrassment last night	37	0	32.87
fresh emerged it that	34	0	30.2
fresh a facing the	33	0	29.31
fresh facing last night	33	0	29.31
fresh a in row	30	0	26.65
fresh a in over	30	0	26.65
fresh a of row	29	0	25.76
fresh in night the	28	0	24.87
fresh of row the	28	0	24.87
fresh a of today	28	0	24.87

anew functions in a way similar to *fresh* as it highlights the news value of recency and stresses there is a new development that needs attention.

4.2.2 *Text-initial concgrams with origin* ***row***

The following concgrams built around the origin word *row* are key for their frequency in text-initial sentences:

> *row a embroiled in night* (14-0)*, row broken has out* (25-0)*, row over the yesterday* (196-3), *row after last night the* (81-0), *row a political yesterday* (21-0), *row fresh over yesterday* (18-0)

(9) The government was last **night embroiled in a** dirty tricks **row** after an official memo alleged that **a** minister wanted civil servants to dig up damaging information to "rubbish" the radical Channel 4 comedian Mark Thomas.

The example of *row* (9) contains a verb that we found for the *fresh* pattern (*embroiled in*). The main difference here is that instead of *fresh* we find *dirty tricks* as modifier for *row.* However, both modifiers fulfil a similar function in that they emphasize news values, with *dirty tricks* focusing more directly on the negativity than *fresh.*

4.2.3 *Text-initial concgrams with origin* ***attempt***

The following concgrams built around the origin word *attempt* are key for their frequency in text-initial sentences:

> *attempt blair to tony will* (27-3)*, attempt an by government in the* (27-15), *attempt an court to yesterday* (42-0), *attempt failed yesterday* (38-0), *attempt a ditch last* (5-4)*, attempt block court* (9-0)

(10) **Tony Blair will** today **attempt to** lower tensions in the worst industrial dispute of his premiership when he reassures Britain's 50,000 firefighters that he has no intention of apeing Margaret Thatcher's bitter year-long dispute with the miners.

(11) **Tony Blair will** today hold the first of a series of meetings with his chancellor, Gordon Brown, and the health secretary, Alan Milburn, in an **attempt to** broker a compromise over plans **to** create free-standing foundation hospitals in the NHS – enjoying new financial freedoms on terms that satisfy both men.

The noun attempt is an origin that is not at first sight linked to our original findings for the nucleus pattern. The examples suggest, however, that concgrams of *attempt* are potential candidates to investigate the pattern further. In examples (10) and (11) the time reference is to *today* and the article refers to Tony Blair's attempts to deal with difficult situations, i.e. lowering tensions or brokering a compromise. Thus the text-initial sentence highlights news value in a similar way to the *fresh* pattern.

5. Conclusions

In our Guardian newspaper corpus we have found many items with a strong association with text position (especially text-initial position) when defined in terms of sentence location within paragraph and text. We have adapted the widely used key word (or key-item analysis) procedure to carry out intra-textual comparisons of individual words and clusters (or n-grams). The matrix method provides a systematic and comprehensive account of the text-positional associations of words and clusters. The resulting associations are particularly complex and involve levels of nesting or combination. In previous analyses the contextual examination of a number of words and clusters with text-initial associations revealed frequent co-occurrence within the same sentence, which we interpreted as instances of a nucleus pattern to introduce the main meanings of an article and highlight its news values. While the initial discovery of the nucleus pattern has been a time consuming process the extraction of concgrams (flexible sets of co-occurring items in specific contexts, e.g. the sentence) can

be used along with key-item analysis to extract co-occurrence patterns with particular text-positional associations. We used the WSConcGram program in the latest version of WordSmith Tools to test this notion. Concgrams can be seen as candidate sets of words for more detailed textual analysis. Our analysis illustrates how methods from corpus-linguistics, when targeted to specific textual positions, can complement text-linguistic analyses. Specifically, our findings relate to the work of White (1997) on the structure of news text. Words, clusters and concgrams that tend to occur in the first sentence of a newspaper text form part of the core/nucleus of the story and play a central role in the structure and organization of the text.

Appendix

Analysis of a Home News article from *The Guardian* (11 January 2000, David Hencke) following the 'orbital structure' proposal for hard news stories as presented in White (1997).

Nucleus (headline and first sentence)

Backlash over Straw veto on release of information

The home secretary, Jack Straw, waded into a fresh bout of controversy over his much-criticised freedom of information bill yesterday when it emerged he had given himself a new veto to ban the release of any public documents.

Specification 1:
The proposal, which was put into the bill without consulting Parliament, united in opposition Labour rebel MPs, Liberal Democrats and Conservatives on the eve of the measures facing scrutiny in the Commons. The bill was already certain to face a rough ride through Parliament because Mr Straw had blocked reforms to allow the new information commissioner power to override ministers to order publication of contentious information and insisted on wide-ranging exclusions to the release of documents, particularly in policy-making areas.

Specification 2:
Mark Fisher, the former arts minister sacked by Tony Blair in his first reshuffle, said yesterday: "This proposal is an extraordinarily huge loophole. It basically allows Jack Straw to put a razor through the whole bill by removing access to any piece of information from the public domain by parliamentary order."

Specification 3:
The loophole was only spotted by campaigners and MPs last week when they realised the home secretary had inserted the new powers when he revised the

bill before Christmas. The clause says the home secretary may by order limit the release of information by any government department, the Houses of Parliament, Northern Ireland assembly, the Welsh assembly and the armed forces.

Specification 4:
Maurice Frankel, director of the Campaign for Freedom of Information, said: "This is an outrageous measure which could allow Jack Straw in theory to reduce the information by any government to just a public inquiry number. It is also giving the same power to any successor government."

Specification 5:
MPs are particularly angry that Mr Straw is proposing to use an arcane parliamentary procedure called "a negative resolution" to obtain these powers. This will allow him to ban release of any particular piece of information through an obscure parliamentary order which will become law if it is not spotted by the opposition within 28 days.

Specification 6:
The proposals have angered the Liberal Democrats and Robert Maclennan, MP for Caithness, Sutherland and Easter Ross, and David Heath, MP for Somerton and Frome, have tabled an amendment deleting the proposal.

Mr Maclennan said last night: "This amounts to a dispensing clause. It basically says we shall apply the law, except when Jack Straw says so. This seems to be based on excessive caution."

Specification 7:
Ann Widdecombe, the shadow home secretary, has also sought to ban the move apart from in cases when a block can be said to be against the public interest.

Specification 8:
The home office claimed the new clause was necessary to modernise government legislation. A spokeswoman said: "If a future home secretary used this clause to try to wipe out large sections of the bill I don't think the parliamentary authorities would stand for it."

Bionotes

Matthew Brook O'Donnell is a research fellow in the English Language Institute, University of Michigan and involved in projects including MICASE and MICUSP, contributing expertise in corpus compilation, annotation and the development of computational tools for analysis. His research interests include the integration of corpus and text-linguistic methods, the study of language

acquisition in terms of lexical associations and usage-based theories, as well as the application of techniques from machine learning and natural language processing to corpus linguistic tools and methods. Email: mbod@umich.edu

Mike Scott, after many years in EFL and ESP working in Brazil, Mexico and at Liverpool University, now a researcher with Aston University, spends most of his time developing his corpus linguistic software, WordSmith Tools. Email: mike@lexically.net

Michaela Mahlberg is Associate Professor in English Language and Applied Linguistics at the University of Nottingham. She is the author of *English general nouns: A corpus theoretical approach* (John Benjamins, 2005), she published the book *Text, discourse and corpora. Theory and analysis* (Continuum 2007, jointly with Michael Hoey, Michael Stubbs and Wolfgang Teubert). She is the editor of the *International Journal of Corpus Linguistics* (John Benjamins), and co-editor of the series *Corpus and Discourse* (Continuum). Email: Michaela.Mahlberg@nottingham.ac.uk

Michael Hoey has been Baines Professor of English Language at the University of Liverpool since 1993 and is currently Pro-Vice-Chancellor for Internationalisation, and Director of the University of Liverpool's Confucius Institute. He has published over 80 articles, is the author of six books and is the editor of two. One of his books (*Patterns of Lexis in Text,* 1991) won the Duke of Edinburgh English Speaking Union Award for best book in applied linguistics. He is co-editor with Tony McEnery of a series of books on advances in corpus linguistics. He has lectured in over 40 countries. Email: hoeymp@liverpool.ac.uk

Notes

1. Römer (2010) includes text-colligation as one of the four steps in her Phraseological Profile (PP) method that is applied to a specific corpus made up of book reviews from the discipline of linguistics.
2. We are grateful to the AHRC for their support (AHRC grant Ref. 119390) that has made this research possible. Further results from the project are presented, for instance, in Hoey and O'Donnell (2007a, 2007b, 2008a, 2008b, 2009).
3. Otherwise known as n-grams or lexical bundles. In this study we use the terms clusters or n-grams. But to avoid confusion we do not use the term *lexical bundle* as these are defined according to specific frequency thresholds and dispersion measures (Biber et al. 2004). In collecting clusters we made use of the raw frequency threshold of 5 set up in WordSmith Tools.
4. David Hencke, 'Backlash over Straw veto on release of information' *The Guardian*, Tuesday 11 January 2000 (available http://www.guardian.co.uk/politics/2000/jan/11/freedomofinformation.uk – December 2009).
5. The default settings in the KeyWords tool of WordSmith Tools 5.0 were used for the key-item analysis. These make use of the log-likelihood measure with a significance at the level $p < 0.000001$.

6. In other studies we have used just one comparison to define text-initial (e.g. Hoey and O'Donnell 2008b, Forthcoming, where text-initial items are taken from a TISC against NISC comparison) or grouped the sub-corpora differently (e.g. Mahlberg and O'Donnell 2008, where text-initial is defined by comparing TISC against all other sub-corpora, termed nTISC).
7. See Mahlberg (2009) for an analysis of the pattern *move follow** in the first sentence of the second paragraph of Guardian articles and its textual function in linking back to the nucleus of the article.
8. It is beyond the scope of this paper to discuss the psychological reality (or otherwise) of collocation and other associations posited in corpus linguistics. Lexical Priming theory states these patterns, which are observable in large language corpora, in psycholinguistic terms without extensive reference to experimental data or psycholinguistic literature. Hoey (2010) attempts to address this issue and a number of recent experimental studies do offer support to the claims (see, for example, Ellis et al. 2009; Ellis and Frey 2009; papers in *Corpus Linguistics and Linguistic Theory* 2009 Volume 1).
9. See Hoey and O'Donnell (Forthcoming) for a demonstration of this for the cluster *according to a*. The examples in Section 2.2 and 2.3 serve to illustrate the application of the macro-analysis of a corpus in terms of textual structure to the micro-level of an individual text. However, there are many implications of how this relates to text linguistic and discourse analytic theories of structure and organization. Some of these are worked out in Hoey and O'Donnell 2007a, 2007b, 2008a, 2008b and 2009.
10. Mahlberg (2009) relates her findings for the noun *move* to the theory of orbital structure of hard news texts (White 1997) and her own theory of Local Textual Functions (Mahlberg 2005, 2007).
11. In his software demonstration on concgrams at the conference in Siena on Keyness, Chris Greaves (2007), who develops ConcGram (Greaves 2009), stated that in order to extract the four and five concgrams from a corpus of 5 million words, his team had to split the analysis across nearly 100 computers over a period of weeks. Greaves (2009: 12) states that ConcGram 'is intended as a tool for text analysis, and automatic searches are best conducted on files which do not have more than 5 million words, and are faster on smaller files. For example, a corpus of about 1 million words with 18,000 unique words will take about a day to create the initial 2-Word Concgram List on a PC running Windows XP with a Pentium 4 3 GHz CPU and 2 gigabytes RAM'. The design of WSConcGram should in principle allow corpora of 100 million words (e.g. the BNC) to be processed. At the current stage of development it has been successfully tested on around 10–12 million words of text.
12. In calculating key words and key clusters the relative frequency of each item in the target text or corpus and the reference corpus is calculated using the token size of each of these respectively (see Scott 1997; Scott and Tribble 2006). It is not immediately clear how best to normalize concgram frequencies (tokens) such as those in Table 6. In the key concgram comparison here we use the total number of concgrams for the origin word (fresh) found by WSConcgram in TISC and NISC. Another possibility is to use the total number of single items found by WSConcgram for each corpus (13,473 and 20,226 respectively).

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Ulla Connor, & Thomas A. Upton. 2007. *Discourse on the move. Using corpus analysis to describe discourse structure*. Amsterdam/Philadelphia: John Benjamins.

Biber, Douglas, Susan Conrad, & Viviana Cortes. 2004. If you look at . . . : Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25. 371–405.

Baker, Paul. 2006. *Using corpora in discourse analysis*. London: Continuum.

Cheng, Winnie, Chris Greaves & Martin Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4). 411–433.

Cheng, Winnie, Chris Greaves, John Sinclair & Martin Warren. 2009. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics* 30(2). 236–252.

Csomay, Eniko & Viviana Cortes. 2010. Lexical bundle distribution in university lectures. In Stefan Th. Gries, Stefanie Wulff & Mark Davies (eds.), *Corpus linguistic applications: Current studies, new directions*, 153–168. Amsterdam: Rodopi.

Ellis, Nick C., Eric Frey & Issac Jalkanen. 2009. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In Römer, Ute & Rainer Schulze (eds.), *Exploring the lexis-grammar interface. Studies in corpus linguistics*, 89–114. Amsterdam: John Benjamins.

Ellis, Nick C. & Eric Frey. 2009. The psycholinguistic reality of collocation and semantic prosody: Affective priming. In Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali & Kathleen M. Wheatley (eds.), *Formulaic language: Volume 2. Acquisition, loss, psychological reality, and functional explanations,* 473–498. Amsterdam: John Benjamins.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. University of Stuttgart, Stuttgart.

Francis, Gill. 1994. Labelling discourse: An aspect of nominal-group lexical cohesion. In Malcolm Coulthard (ed.), *Advances in written text analysis*, 83–101. London: Routledge.

Granger, Sylviane & Fannie Meunier (eds.). 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Greaves, Chris. 2007. *Software demo: ConcGram*. Conference on Keyness in Text, Siena. http://www.disas.unisi.it/keyness/index.php (accessed December 2009).

Greaves, Chris. 2009. *ConcGram 1.0. A phraseological search engine*. Amsterdam: John Benjamins.

Hoey, Michael. 2004. Lexical priming and the properties of text. In Partington, Alan, John Morley & Louann Haarman (eds.), *Corpora and discourse*, 386–412. Bern: Peter Lang.

Hoey, Michael. 2005. *Lexical Priming: A new theory of words and language.* London: Routledge.

Hoey, Michael. 2010. Semantic Priming. In Patrick C. Hogan (ed.), *The Cambridge encyclopedia of the language sciences*. Cambridge: Cambridge University Press.

Hoey, Michael & Matthew Brook O'Donnell. 2007a. Death to the topic sentence: How we really paragraph. In Leung, Y. N. (ed.), *Selected papers from the Sixteenth International Symposium on English Teaching*, 60–76. Taipei: English Teachers' Association/ROC.

Hoey, Michael & Matthew Brook O'Donnell. 2007b. How we really paragraph. In G. Śpiewak, *The Teacher* (Macmillan Teaching Adults Series, December 2007).

Hoey, Michael & Matthew Brook O'Donnell. 2008a. Lexicography, grammar, and textual position. *International Journal of Lexicography* 21(3). 293–309.

Hoey, Michael & Matthew Brook O'Donnell. 2008b. The beginning of something important?: Corpus evidence on the text beginnings of hard news stories. In Barbara Lewandowska-Tomaszczyk (ed.), *PALC 2007 Practical Applications In Language Corpora: 7* (Studies in Language), 189–212. New York: Peter Lang.

Hoey, Michael & Matthew Brook O'Donnell. 2009. The Chunking of Newspaper Text. In M. Shiro, P. Bentivoglio & F. Erlich (eds.), *Haciendo discurso. Homenaje a Adriana Bolívar*, 433–452. Comisión de Estudios de Postgrado de la Facultad de Humanidades y Educación de la Universidad Central de Venezuela.

Hoey, Michael & Matthew Brook O'Donnell. Forthcoming. 'According to a study by . . .'. *International Journal of Corpus Linguistics*.

Mahlberg, Michaela. 2005. *English general nouns: A corpus theoretical approach*. Amsterdam: John Benjamins.

Mahlberg, Michaela. 2007. Lexical items in discourse: Identifying local textual functions of *sustainable development*. In Michael Hoey, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert. *Text, discourse and corpora*. *Theory and analysis*, 191–218. London: Continuum.

Mahlberg, Michaela. 2009. Local textual functions of *move* in newspaper story patterns. In Römer, Ute & Rainer Schulze (eds.). *Exploring the lexis-grammar interface. Studies in corpus linguistics*, 265–287. Amsterdam: John Benjamins.

Mahlberg, Michaela & Matthew Brook O'Donnell. 2008. A *fresh* view of the structure of hard news stories. In Stella Neumann & Erich Steiner (eds.) *Online Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop, Saarbrücken, 23–25 July 2007*. http://scidok.sulb.uni-saarland.de/volltexte/2008/1700/.

O'Donnell, Matthew Brook & Ute Römer. In preparation. Positional variation of n-grams and phrase-frames in a new corpus of student writing.

Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1). 95–119.

Römer, Ute & Rainer Schulze. 2009. Patterns in language: An introduction. In Reinfandt, Christoph & Lars Eckstein (eds.), *Anglistentag 2008 Tübingen. Proceedings.* 359–365. Trier: Wissenschaftlicher Verlag Trier.

Scott, Mike. 1997. PC analysis of key words – and key key words. *System* 25(2). 233–45.

Scott, Mike. 2002. Picturing the key words of a very large corpus and their lexical upshots or getting the Guardian's view of the world. In Bernard Ketteman & Georg Marko (eds.), *Teaching and learning by doing corpus analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, July, 2000*. Amsterdam: Rodopi.

Scott, Mike. 2008. *WordSmith Tools Version 5.0*. Lexical Analysis Software, Liverpool.

Scott, Mike & Chris Tribble. 2006. *Textual Patterns: Key words and corpus analysis in language education.* Amsterdam: John Benjamins.

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

Sinclair, John, Susan Jones & Robert Daley. [1970] 2004. *English collocation studies: The OSTI report*. London & New York: Continuum.

Stubbs, Michael. 1996. *Text and corpus analysis*. Oxford: Blackwell.

White, Peter. 1997. Death, disruption and the moral order: the narrative impulse in mass-media 'hard news' reporting. In Frances Christie & J. R. Martin (eds.), *Genre and institutions. Social processes in the workplace and school*, 101–133. London: Continuum.

Examples quoted from the *Guardian*:

(1)–(4) and full article in the Appendix: 'Backlash over Straw veto on release of information', 11 January 2000, David Hencke, p. 1

(5) 'Woodland pay fuels "fat cat" row', 14 September 1998, Rebecca Smithers and John Carvel, p. 6

(6) 'Labour in trouble over gifts', 13 August 2003, Nicholas Watt, p. 10

(7) 'Tories protest at history shift', 5 August 1999, Rebecca Smithers, p. 4

(8) 'Lottery falls pounds 5bn short', 28 June 2001, Matt Wells, p. 3

(9) 'Whitehall tried to smear comedian', 8 January 2001, Rob Evans and David Hencke, p. 3

(10) 'Blair takes reins in fire dispute', 25 November 2002, Nicholas Watt, Kevin Maguire, David Gow and Larry Elliott, p. 1

(11) 'Blair tries to heal cabinet rift over hospitals', 7 October 2002, Michael White, p. 12