# Key colligation analysis: Discovering stylistic differences in significant lexico-grammatical units

Nozomi Miki
Kansai University of International Studies
Nara University of Education

**Abstract**

This methodology-based research proposes key colligation analysis to deals with grammatical relations in a particular dataset at a phrasal level based on statistics, while keyword analysis treats statistically outstanding words rather than phrases, irrespective of grammatical relations or the word class. The method can reveal naturally occurring sequences of subjects and predicates, including verbs, voice, and auxiliaries, if any, so as to detect the subtle stylistic differences which they create between the texts. This is demonstrated by British newspaper editorials, which share the same register and genre but bear distinct characters. The results reveal a diversity of argumentative styles: personality vs. impersonality, involvement vs. detachment, authority vs. intertextuality, for instance. Since quality newspapers are a precious resource for research, it is worthwhile investigating the variations among them for proper linguistic generalisation. The implications of this research are applicable to disciplinary variations.

**Key words**:

key colligation analysis, subjects and predicates, British quality newspapers, editorials

## 1. Introduction

Grammatical subjects are not always but very often themes—'a departure of message' in the clause-initial position (Halliday, 1967: 212, see also Fries, 1981, 1983, 1994). The investigation of subject nouns shows us how writers take their positions. In argumentative writing in particular, selecting writers themselves as subject, or a theme for the clause-initial position, often has a great impact on textual interaction (Francis, 1989: 201; Gosden, 1993: 57; Martin 1985a; 41), where writers position themselves in relation to readers and also stand in relation to the argument, as Hyland (2002) indicates:

> While the important focus of academic writing tends to be the events or concepts under discussion in the rheme, ***the choice of first position is very significant***. The way a writer begins a clause not only foregrounds important information, firmly identifying the writer as the source of the associated statement, but also helps the writer control the social interaction in the text (e.g., Gosden, 1993) (Hyland, 2002: 1093, boldfaced by the author)

This tendency is obvious in the use of first person pronouns, which mark the overt presence of an author, that is, authority. The manifestation of such authority denotes writers' active interaction in the text with readers (Hyland, 2001: 218). In fact, several studies have revealed a

variety of discourse functions for first person pronouns (Fortanet, 2004; Kuo, 1999; Íñigo-Mora, 2004; Ventola, 1994).

Importantly, it has been reported that subjects show categorical features which mirror not only registers and genres but also sub-genres; Hyland (2001: 212) finds a disciplinary variation of first personal pronouns. Similarly, Hawes and Thomas (1996) indicate the different types of subject which were preferred by a quality newspaper and a tabloid newspaper in the UK (e.g., *The Times* and *The Sun*). Martin (1985a) suggests that subjects play a significant role in determining stylistic tones in expositions. Gosden (1993) discovers lexical subjects as well as the pronominal ones unique to sections of academic writing. Moreover, Charles (2006) and Gosden (1993) make a point of indicating that verbs together with subjects show variations among academic disciplines and sections of academic papers.

However, the efficiency and effectiveness of the methodology to uncover statistically outstanding combinations of subjects and predicates in a particular dataset leave something to be desired. Keyword analysis (Scott, 2000, 2001; Scott and Tribble, 2006), the most widely practiced corpus linguistic method between texts, does not serve this purpose. The way to retrieve subjects together with their predicates has mostly been manual, making investigations so time-consuming as to limit the number of texts being investigated and thus making generalisation hazardous. Otherwise, the main targets were pronominal subjects instead of lexical ones, in quantitative research above all.

This research proposes key colligation analysis, an innovative way to identify particular colligations in the massive number of datasets, where otherwise manual, time-consuming search would usually have to be carried out in a limited number of texts. Colligations, a word which was coined by Firth (1957 [1951]), are significant word combinations with special preference for certain positions to occur in. See the definition below by Hoey (2005: 42–43) (cf. Hunston, 2001)

1. The grammatical company a word or word sequence keeps (or avoids keeping) either within its own group or at a higher rank;
2. The grammatical functions preferred or avoided by the group in which the word or word sequence participates;
3. The place in a sequence that a word or word sequence prefers (or avoids).

(Hoey, 2005: 42–43)

Distinguishing colligations from collocations, recent research proposes key colligation analysis, and then shows how effective and efficient this methodology is in revealing the actual uses of subjects and predicates, on the basis of a massive number of data from editorials.

Quality newspapers are precious resources of studies of English usage and editorials reflect national writing in argumentation (Connor, 1996: 143), since a number of contrastive studies employ this type of text. The investigation of uses of subjects in editorials offers ideal samples for novice and L2 writers who find it less than easy to choose right subjects in the right place in persuasive prose (for the overuse of the first person pronouns by L2 writers, McCrostie 2008; Petch-Tyson, 1998; Ringbom, 1998; for ‘writers’ identification, Baynham, 1999; Cadman, 1997; Hyland, 2002; Ivanič, 1994). Above all, since we deal with quality newspapers, our way can be applied to the investigation of disciplinary differences.

## 2. Previous studies

### 2.1 Themes and subjects

Whether clause-initial elements are themes is arguable. Some researchers relate theme to old/new information, while one relates it to meanings (Halliday, 1985) and others believe that this concept should be determined in terms of texts as well as meanings (Fries, 1983). Recent researchers of genres have used themes as a catch-all phrase including grammatical subjects (cf. Hawes and Thomas, 1996: 159–160); specifically, subjects are regarded as unmarked themes (Martin, 1986b) or a realisation of 'obligatory topic', in contrast to adjuncts as 'optional context frame,' as in "*In this paper*, we describe …"; Gosden (1993: 58–60) calls it 'obligatory,' on the basis of the ratio of nominal groups of themes in his research. The present paper adopts a separation of subjects from other thematic, clause-initial adjuncts, for two reasons. First, themes themselves take different shapes, ranging from adverbs to clauses. Such a wide range makes it difficult to focus and to carry out qualitative analysis. Secondly, though I could not agree more with the view that adjuncts also play a 'pivotal role' in textual organisation so as to guide readers (Gosden, 1993: 59), subject nouns play a significant role in determining stylistic atmosphere.

Subject nouns mirror distinct registers. Fries (1983: 121) indicates that 'the point of departure of each clause is related to the previous context in argumentative or expository prose. This finding was developed by Francis (1989) and Hawes and Thomas (1996). Francis (1989) mentions that themes of news reporting were about 'what people say,' leading to more verbal characteristics, while those of editorials and letters to editors were more relational, in terms of the use of nominalisation.

Subjects can also be indicators of stylistic variations within the same genres; editorial styles in different newspaper companies, for instance. Martin (1985a) proposes two divisions of argumentative writing: Hortatory Expositions and Analytical Expositions. In Hortatory Expositions, human subjects are highly visible, together with verbs of perception, feeling, thinking and saying in the active voice. In contrast to such an interactive style, Analytical Expositions create an information-giving tone, as seen in their impersonal constructions, verbs of saying whose subjects are inanimate such as data and reports, and the passive. Hawes and Thomas (1996) point out that a quality British newspaper, *The Times,* exhibited a high frequency of institutional subjects, in contrast to a tabloid, *The Sun*, which revealed more instances of human subjects, such as *we* and individual names. Westin (2002), who conducted chronological research on stylistic variations of British newspaper editorials throughout the 20th century, finds that *The Guardian* and *The Daily Telegraph* displayed more instances of first person plurals than *The Times*, which is regarded as a piece of evidence that *The Guardian* was relatively informal.

These studies are insightful but the number of data and statistics used were not enough to allow generalisation. Martin (1985a) is no more than a pilot study of two texts from two magazines in different countries. Hawes and Thomas (1996) do not release the tokens but the normalised frequencies in spite of the different lengths of tabloids and quality newspapers, thus denying statistical validity to the findings. Normalised frequencies are likely to distort the number, making it look larger than they should. The newspaper corpora in Westin (2002) were not very large, considering the period of time (100 years); each newspaper corpus came to 0.5 million words in total or 15,000 to 19,000 words per year. This possibly affects the results; in fact, my previous research, based on larger datasets (about 300,000 words in each newspaper corpus), finds that *The Guardian,* like *The Times,* did not display a higher frequency of first person plurals than *The Daily Telegraph* (Miki, 2010b), though I conducted the analysis according to his mode of investigation (i.e., excluding the first person plurals in quotations). Westin regards *The Guardian* as informal, basing his view on the total of this kind of pronoun

in comparison with that in other newspapers, though he notes the decreasing number of first person plurals in *The Guardian* during the 20th century, which is attested by the Newman-Keuls multiple comparison test. The synchronic study of stylistic variations of editorials is necessary with a massive dataset and appropriate statistics.

**2.2 Reconsidered keyword analysis and follow-up analysis**

In order to retrieve statistically the more outstanding words in a corpus, keyword analysis is widely practiced. This statistical comparison, based on the log-likelihood ratio, tells us the aboutness and styles of the texts of interest. However, this method does not serve our purpose of seeing subjects unique to a particular data collection. A keyword list is a miscellaneous set of words in terms of grammar, word-forms, and functions. See the keyword list of the editorials from *The Guardian* (Figure 2.1).

File Edit View Compute Settings Windows Help

| N | Key word | Freq. | % | RC. Freq. | RC. % | Keyness | P | emmas | Set |
|---|---|---|---|---|---|---|---|---|---|
| 1 | GUARDIAN | 419 | 0.02 | 69 | | 769.32 | 0.0000000000 | | |
| 2 | BIG | 1,236 | 0.06 | 1,355 | 0.02 | 553.15 | 0.0000000000 | | |
| 3 | AL-QAIDA | 205 | 0.01 | 2 | | 529.21 | 0.0000000000 | | |
| 4 | BUT | 14,891 | 0.73 | 33,592 | 0.58 | 516.91 | 0.0000000000 | | |
| 5 | WHICH | 7,485 | 0.37 | 15,620 | 0.27 | 452.63 | 0.0000000000 | | |
| 6 | YESTERDAY'S | 994 | 0.05 | 1,202 | 0.02 | 372.43 | 0.0000000000 | | |
| 7 | HIZBULLAH | 131 | | 0 | | 351.80 | 0.0000000000 | | |
| 8 | THINGS | 992 | 0.05 | 1,237 | 0.02 | 349.74 | 0.0000000000 | | |
| 9 | YESTERDAY | 3,399 | 0.17 | 6,428 | 0.11 | 344.92 | 0.0000000000 | | |
| 10 | LAST | 3,431 | 0.17 | 6,554 | 0.11 | 332.42 | 0.0000000000 | | |
| 11 | JUST | 2,516 | 0.12 | 4,511 | 0.08 | 320.07 | 0.0000000000 | | |
| 12 | GOT | 532 | 0.03 | 489 | | 311.09 | 0.0000000000 | | |
| 13 | ONE | 5,680 | 0.28 | 12,269 | 0.21 | 273.82 | 0.0000000000 | | |
| 14 | ICM | 127 | | 11 | | 270.97 | 0.0000000000 | | |
| 15 | CURRENT | 909 | 0.04 | 1,234 | 0.02 | 266.72 | 0.0000000000 | | |

Figure 2.1: Keyword list of editorials in *The Guardian*

In the keyword list there are culture-specific and/or topic-specific words, such as *Guardian* and *Al-qaida*, while it also includes functional words of argumentation, such as *but*. However, it is so mingled that it is time-consuming to sort out and find what corpus linguists are interested in, unless the number is limited. In order to categorise features among keywords, researchers need to cope with an overwhelming number of keywords or usually look at the top keywords only, despite the same significance, leaving the possibility of categorical features buried under the surface.

The recent corpus studies then have put forth ways to target keywords with certain features. Key categories proposed by Baker (2004, 2006) offer a way of retrieving statistically outstanding words with specific POS tags and semantic annotations in a particular dataset. Key associates (Scott and Tribble, 2006) cover a set of keywords with a node keyword across a range of texts (i.e., a key keyword) so as to see the words associated with a particular word. Key clusters (Baker, 2004, 2006) reveal statistically more frequent n-grams in a target corpus than in a normed corpus. Although there are various applications of keyword analysis, none of them treats key sequences with grammatical relations. For instance, key clusters analysis (Baker, 2006: 140–141) retrieves sequences at random, so that clusters may be frequent bits of grammatical units (e.g. *often meet the*).

The 'key' analyses are instances of quantitative investigation, while the previous discourse studies suggest that subjects and predicates collaboratively work in textual interactions with readers. With the trade-off relation between data sizes and qualitative analysis, discourse analysis treats lexical subjects in a small dataset, while quantitative analysis begins with a

closed category such as pronominal subjects and examines the co-occurring verbs, or auxiliaries, or beings with a particular type of verbs such as verbs of attribution and then subject nouns. It is certainly worth investigating subjects and predicates together rather than separately, keeping them as natural or as close to primary sequences as possible; however, few quantitative studies have focused on such strings, which are based on sufficient datasets and statistics.

Here, Goto (2006, 2008) is an exception; he retrieved subject-verb combinations from the BNC academic sections, using *Machinese Syntax*, a parser, which catches subjects and verbs. Although Goto (2006) used raw frequencies and checked the distribution of all words with correspondence analysis, his next work (Goto, 2008) employed statistical tests to retrieve specialised grammatical units, based on the complimentary similarity measure (CSM). Since his aim was to make lists of lexical subject-verb pairs for EAP practitioners, his research excluded expletive subjects such as *there* and *it* and also the passive and auxiliaries. However, the excluded elements can play a significant role in argumentation. Actually, the list of frequent lexical subjects and verbs without expletives, passives and auxiliaries does not reflect actual usage and linguistic reality and thus fall short of revealing stylistic variations. His programming script also suggests that subjects and verbs in the main clauses were not distinguished from those in the subordinate clauses, though this is crucial in evaluation studies of attribution (or intertextuality) and averral.

Even so, Goto's idea of retrieving units of subjects and verbs together with their frequency is notable and applicable to the present research. This seeks to retrieve subjects and predicates in much the same way as his. At the same time, this research covers pronominal subjects, modal auxiliaries, and the active/voice information, all of which are concerned with the interpretation of subject and verbs, and employs a different statistical index. The lexico-grammatical units as such are not merely frequent word combinations, or collocations, but 'colligations.'

## 3. Methodology: Key colligation analysis

Key colligation analysis begins by parsing texts or marking words with grammatical categories. For this purpose, *Machinese Syntax* (Connexor Oy, 2008) was chosen, where the way of parsing texts is based on Functional Dependency Grammar (FDG) (see Tapanainen, 1999); one word depends on another and on the basis of this dependency. However, the verbs in the main clause are exempt from this dependency; each sentence has a single main verb, which means that main and subordinate verbs are marked differently. The main verb is independently located as a root or a starting point of every dependency.

One of the strengths of this software tool lies in its value for structural analysis, which is not affected by surface word orders as long as the syntactic functions of words remain the same (Tapanainen, 1999: 18). In collocation analysis, whose span is usually within three words before or after a node word, it is if not impossible, it is usually quite difficult to capture them as a unit; with longer spans, the manual search and follow-up analysis become more time-consuming. Let us take (1), for example, where a subject, *newspaper* is not next to a main verb, *argued*:

(1) This ***newspaper*** has for some time ***argued*** for more democratic control over local constabularies—an aim to which all three parties have signed up, with varying degrees of sincerity. (*The Daily Telegraph*, April 13, 2004, "Controlling the police", boldfaced by the author)

The subject and the verb are separated by an aspect, *have* and an adverbial phrase, *for some time*. However, they are appropriately parsed with *Machinese Syntax*, as seen in Table 3.1:

Table 3.1: Distant colligations and *Machinese Syntax*

| # | Text | Base form | Syntactic relation | Syntax and morphology |
|---|---|---|---|---|
| 1 | this | This | det:>2 | @DN> %>N DET DEM SG |
| 2 | **newspaper** | newspaper | **subj:>3** | @SUBJ %NH N NOM SG |
| 3 | has | Have | v-ch:>7 | @+FAUXV %AUX V PRES SG3 |
| 4 | for | For | dur:>7 | @ADVL %EH PREP |
| 5 | some | Some | det:>6 | @DN> %>N DET |
| 6 | time | Time | pcomp:>4 | @<P %NH N NOM SG |
| 7 | **argued** | Argue | **main:>0** | @-FMAINV %VA EN |
| 8 | for | For | phr:>7 | @ADVL %EH PREP |
| 9 | more | Much | qn:>11 | @QN> %>N DET CMP SG |
| 10 | democratic | democratic | attr:>11 | @A> %>N A ABS |
| 11 | control | Control | pcomp:>8 | @<P %NH N NOM SG |

Note. This is a part of the parsed sentence.

In the above table, the base forms and the syntactic relations represent lemmatisation and grammatical and thematic relations respectively; the grammatical subject is marked by 'subj,' while the main verb is referred to as 'main.' It should be remembered that, since the root of every dependency is main:>0 in the syntactic relation, @SUBJ in the column of syntax and morphology is additional information, which comes after the syntactic relation. Retrieving subjects from parsed texts should be based on as much information about a syntactic relation as possible; and this makes a difference. In fact, the subject information in syntax and morphology does not distinguish the main clauses from the subordinate ones. We used the information of syntax and morphology, so as to narrow subjects down to a particular type; expletives, for instance. Then the numbers after the >, greater-than signs are the numbers in the # column, which indicate the forms of linkage between words, or the dependency. They are the determinant elements of the grammatical relations. For instance, the subject status of *newspaper* depends on the third element, *has*; the v-ch status of *has* is dependent on the fourth element, *argued*.

In addition, the lemmatisation of words makes *Machinese Syntax* more precise and recoverable than other similar parsers do. Goto (2005), using test data in Hindle (1994), showed that *Machinese Syntax* marked a higher score in the precision of parsing verbs (95%) and object nouns and the recall of them (83% ) than *Link Parser*; according to him, this can be attributed to lemmatisation before parsing. Although *Machinese Syntax* is a better parser, the fact is that no parsers yet mark a full score of precision and recall, so I checked the significant colligations of my interest manually.

The current research sets up several policies in retrieving colligations from the texts parsed with *Machinese Syntax*. First, subjects in the main clause were differentiated from those in the subordinate and relative clauses considering the distinct textual and interpersonal effects of clause-initial subjects and the question of averral and attribution (Hunston, 2000: 178). Second, our analysis covers expletive subjects such as *there* in *there*-constructions, which is differently annotated (with @F-SUBJ in syntax and morphology), as well as lexical ones (marked as @SUBJ). Third, in our investigation, the voice information is kept in retrieving colligations: %AV for the active voice and %PV for the passive voice in the column of syntax and morphology, so that *we were told ...* and *we told ...* can be treated differently (cf. 'transitivity' in Oktar, 2001: 323). Fourth, since modal auxiliaries not only mirror writers' overt attitudes but are also covert rhetorical techniques, those in the texts, if any, were retrieved together with subjects and with other predicates. Lastly, I did not retrieve *be* in the progressive and *have* in the perfect, both of which are marked v-ch:># in the syntactic structure and %AUX in the morphology and syntax. Verbs are lemmatised (e.g., *running* as *run*), so that items with

the same meanings but different forms would not appear as sparse or be counted separately, leading to infrequency in analysis. To sum up, four pieces of information, namely, subj:>#, main:>0, %VA, AUXMOD were retrieved in the investigation.

Below is a summary of our policies to retrieve colligations;

1. Discerning subjects in the main clauses from those in the subordinate clauses
2. Including *there* in *there*-constructions
3. Retrieving modal auxiliaries, if any, together
4. Discriminating the active voice from the passive voice
5. Summarising the tense and the aspect as base forms (i.e., lemmatisation)

From the original sequences or possibly every sequence obtained in the above way, I created four different units and saved them as four lists (see Table 3.2).

Table 3.2: Types of sequences of colligation

| Types of colligation | Actual sequences |
|---|---|
| Subjects | blair, it, we |
| verbs and the voice | say_**a**, say_**p**, remember_**a** |
| subjects, verbs, and the voice | blair_say_**a**, it_say_**p** we_remember_**a** |
| subjects, auxiliaries, verbs, the voice | we_should_remember_**a** |

Key colligation analysis gives us a good focus but also provides us with flexibility; we can switch multi-word units into single-word units at any time, according to the research purposes. For instance, the original sequences in Table 3.2 were used to see the overall tendency of natural combinations, together with their frequencies. The other three were used for a special focus on particular units. Subject-verb sequences were chosen for a sole focus, excluding other elements such as auxiliaries. The frequency of colligations was compared between newspapers, on the basis of statistical findings, which are discussed next.

Log-likelihood was chosen for the statistical comparison of the sequences of grammatical relations. According to Ishikawa (2008), there are two major statistical indices for frequency; the log-likelihood ratio (henceforth, LL or LL ratio) and the chi-square test. These statistical indices are similar, in that the higher the value of $X^2$ or the LL ratio of a word in a particular corpus, the more specialised the word is in the corpus. However, with small corpora, LL is more robust than the chi-square test (Dunning, 1993) and is widely practiced as a way of identifying the vocabulary unique to a particular dataset (cf. Baker, 2004, 2006; Scott, 2001; Tribble, 2000; Scott and Tribble, 2006). In this respect, our choice is reasonable, considering the size of *The Daily Telegraph* compared with the others.

The significance levels and critical values were fixed according to the website of the University of Lancaster Centre for Computer Research (http://ucrel.lancs.ac.uk/llwizard.html). The problem is a choice of cut-offs to limit a daunting number of key items. As Baker (2006) demonstrates in his datasets, changing cut-offs from $p<.000000$ to $p<.001$ means a reduction of 88 keywords to 12. It is necessary to set a cut-off to keep the number manageable; it is virtually impossible to analyse all of them, so most researchers look only at the most significant items. With cut-offs, we can categorically see which units in a target data collection are more specialised than others and are worth further investigation. In linguistics, where language behaves differently from the targets of the social sciences, $p<.05$ is not so rigid when used to characterise the data in question in a reasonable way (Oakes, 1998). $P<.001$ was chosen in the present research. Although I mostly avoided $p<.05$, I used it only when there were fewer key

colligations obtained overall, since the number of subjects which are obtained from a single clause was not particularly large.

One further method is key keyword analysis—keywords with a certain frequency across several files obtained with batch files (Scott, 2010; Scott and Tribble, 2006). Scott (2010) defines a key keyword as "one which is 'key' in more than one of a number of related texts. The more text it is 'key' in, the more 'key key' it is". It sounds reasonable to exclude keywords which occur in few or single texts instead of being dispersed evenly across texts. However, this research did not carry out key keyword analysis, for two reasons. First, the editorial corpora in this research consist of four corpora, which are saved as a single text file comprising more than 3,500 texts each (14,786 in total) for parsing. With the overwhelming number of texts, it is, if not impossible, difficult to deal with the data text by text. Second, Scott and Tribble (2006: 77–84) report that key keyword lists were not different from ordinary wordlists; rather, the genre selection is a key to capturing the representativeness of a target corpus (cf. Baker, 2006; Miki, 2009). The key keyword list from the whole BNC showed that the closed word class (e.g., pronouns and the definite article) was dominant, but far from telling us what the collection of texts was about. Contrastingly, with specific genres, the key keyword lists revealed more of the topic (the aboutness) of the datasets.

Instead of such key keyword analysis (or key key colligation analysis here), I set the minimum raw frequency of key items to five, the same frequency as in key keyword analysis. In fact, the threshold of the raw frequency of a keyword is left to researchers' discretion. The default of the ordinary keyword analysis in *WordSmith* is three, while Scott and Tribble (2006: 72) raise the minimum frequency of three or more in the analysis of the part-of-speech of keywords. We can see the dispersion of key items without using key keyword analysis by searching for them in the concordance and locating the dates and titles of editorial articles. It should be noted that it is not necessary to abandon such recurrent key items if they occur in a couple of texts, or close to that number, following Baker's view of their rhetorical effects. Baker (2006: 143) sends an alarming message to the proponents of key keyword analysis that "… perhaps we should not simply dismiss this particular cluster because its frequency is due to repetition in one speech" (see also Baker, 2004: 350–351).

On the basis of log-likelihood ratios, the key colligations obtained were divided into positive and negative ones, that is, statistically more frequent colligations and less frequent colligations in a target corpus than in a reference corpus. Actually, log-likelihood ratios themselves do not indicate the polarity; at the initial stage they make no difference whether overuse or underuse is found. I added negative and/or positive signs to the LL ratios, according to the normalised frequency; for example, the raw LL ratio of the first person plurals in *The Times* (RF. 503) was 1883.76, exactly the statistical amount without any polarity, in comparison with the reference corpus (RF. 6,377). Given the different sizes of the two corpora, normalisation is applied to determine the opposing directions. In this case, the negative sign is given to the LL ratios of the item in *The Times* (NF. 228.47) in comparison with that in the reference corpus (NF. 1131.12). It is important to note that the LL ratios (cf. critical values) determine the significance of differences in normalised frequency between a target corpus and a normed corpus.

## 4. Data

For the current investigation, a large number of editorials texts over six years were compiled from four leading quality newspapers in the UK. The necessary data were downloaded from *LexisNexis Professionals* and *ProQuest*, web-based search engines which offer access to sources of world news. I checked the list of the titles and the dates with hard copies of newspapers so as to exclude irrelevant articles and removed the headings from the original texts,

while adding the tags of titles and dates to each text and a pair of  and  to each paragraph. See the data sizes in Table 4.1:

Table 4.1: Data specifications

| Target corpora | Tokens | No. of texts | STTR |
|---|---|---|---|
| *The Times* | 2,201,622 | 3,810 | 50.03 |
| *The Guardian* | 2,047,155 | 3,347 | 50.03 |
| *The Daily Telegraph* | 1,569,207 | 3,720 | 50.05 |
| *The Independent* | 2,001,165 | 3,909 | 48.41 |

Each corpus was compared with a reference corpus to produce statistically outstanding sequences. As Baker (2004) and Miki (2009) attest, whether a collection of a variety of English (e.g., BNC or FLOB) or a specialised dataset with the same genre should be employed has a direct impact on the number of keywords. Comparison with the same genres or registers helps to control the number of texts and to reveal subtle or more detailed differences between them. Our reference corpora were tail-ordered from the same genre, editorials, according to the newspapers, so as to locate a style unique to any newspaper in comparison with other newspapers; *The Times* as a target corpus is referenced to the rest of papers as a normed one, for example (see Table 4.2).

Table 4.2: Reference corpora

| Target corpora | Reference corpora | Tokens |
|---|---|---|
| *The Times* | *The Guardian, The Independent, The Daily Telegraph* | 5,485,908 |
| *The Guardian* | *The Times, The Independent, The Daily Telegraph* | 5,830,083 |
| *The Daily Telegraph* | *The Times, The Guardian, The Independent* | 6,166,649 |
| *The Independent* | *The Times, The Guardian, The Daily Telegraph* | 5,700,358 |

It should be remembered that the results are relative; different reference corpora produce a different keyness. The keyness in this article is valid only in comparison with British newspaper editorials.

## 5. Results

This section starts with the overview of key subjects, which are indicators of personality and impersonality, and the key colligations of subjects and predicates, which pinpoint the actual uses. Then it moves to the qualitative analysis such as contextualisation and the investigation of the occurring paragraph (see Table 5.1).

Table 5.1: The top key subjects in the British editorials

| No. | *The Times* | *The Guardian* | *The Daily Telegraph* | *The Independent* |
|---|---|---|---|---|
| 1 | he | that | we | we |
| 2 | it | Labour | Tory | this |
| 3 | policy-maker | # | Howard | there |
| 4 | notion | thing | Letwin | it |
| 5 | per cent | speech | Cameron | fact |
| 6 | Britain | party | reader | truth |
| 7 | airline | byelection | you | government |
| 8 | India | yesterday | Brown | report |
| 9 | China | PCC | newspaper | question |
| 10 | leadership | financier | voter | situation |

Notes. # stands for the total numbers

From the top ten key subjects in *The Times*, the third personal and impersonal singulars are highly visible, followed by abstract nouns and the names of countries. The frequent use of masculine third person singulars suggests that there were a great many politicians in editorials, while the names of countries behave metonymically like a human being, as seen in (2):

(2) … ***China wants*** the US to take the lead in countering terrorism while not being willing itself to take a large political profile. ***It opposed*** the war in Iraq and has continued to withhold support for a more active UN underpinning of Iraq's provisional Government. And ***it has exploited*** Third World resentment of American global power by suggesting that Washington's aims in Iraq and in the Middle East were to establish US political and economic dominance. </p> (*The Times,* October 12, 2004, "China's hostages", boldfaced by the author)

From the above example, *it* does not always acts as an expletive; the pronouns in *it opposed* and *it has exploited* is actually *China*. This suggests that in *The Times* there is a good possibility that impersonal third singulars as subjects are not always expletives but may well be countries. Miki (2011) uses *WordNet* to make a semantic categorisation of subjects, indicating that the greatest variety of country names in this mass medium is among the four quality newspapers. The use of abstract nouns and numbers suggested by *per cent* imparts academic features.

Similarly, *The Guardian* displayed such hard evidence as numbers, which are represented as #, in the top subjects. This paper imparted absolute impersonality, excluding human subjects from the subject nouns. Contrastively, in *The Daily Telegraph*, virtually all the subject nouns were humans; *newspaper* was used in *this newspaper*, which refers to *The Daily Telegraph* itself, that is, the editorial *we*. In *The Independent*, *we* is the most frequent subject but all the rest were non-human, suggesting its middle position in a continuum of personality and impersonality in a different way from *The Times*.

Although the top subject nouns are informative and suggestive about the overall styles, the key colligation list reveals the actual uses of each editorial corpus. First, we look at the overall characteristics on the basis of the positive and negative key colligations, which create an asymmetrical image of what newspapers favoured or disfavoured in the social engagement of readers. Let us begin with the positive and negative key colligations of *The Times* (Table 5.2).

Table 5.2: The top key colligation lists of *The Times*

| Positive key colligations | RF | LL | Negative key colligations | RF | LL |
|---|---|---|---|---|---|
| notion be | 45 | 52.67 | we have | 19 | -194.59 |
| it would be | 603 | 44.93 | we hope | 2 | -97.28 |
| he insist | 60 | 40.78 | we know | 24 | -84.64 |
| issue be | 128 | 33.90 | we need | 21 | -84.60 |
| aim be | 89 | 33.15 | we be | 48 | -82.74 |
| he be | 619 | 31.16 | we see | 6 | -62.70 |
| first be | 184 | 28.38 | that be | 670 | -56.03 |
| he would be | 46 | 28.12 | we hear | 1 | -48.96 |
| pretext be | 11 | 27.94 | fact be | 42 | -39.67 |
| he have | 229 | 26.35 | we learn | 8 | -38.00 |

The mirror image clearly shows that *The Times* tends to avoid first person plurals but not abstract nouns and expletives. In fact, as a key colligation such as *it would be* suggests, *The Times* hit 13 significant *it*-colligations in total, nine of which were accompanied by modal auxiliaries. The modal auxiliaries (e.g., *would, can*) qualify the following evaluative adjectives and nouns in such impersonal patterns: *it would be better/wrong/a mistake/unwise to do …* Another key colligation of *first be* tells us that the subjects in *The Times* play a sequencing role in the text. In the top subject-predicate sequences, the names of countries disappeared, possibly because combinations of countries and predicates were sparse.

Like *The Times*, abstract subjects were dominant in *The Guardian*'s list but not third person singulars, as in Table 5.3**.** This implies the strong non-human tendency of this mass medium in comparison with other newspapers.

Table 5.3: The top key colligations of *The Guardians*

| Positive key colligations | RF | LL | Negative key colligations | RF | LL |
|---|---|---|---|---|---|
| that be | 1,025 | 88.67 | it be | 4,209 | -149.31 |
| thing be | 145 | 49.30 | we have | 36 | -111.50 |
| news be | 86 | 45.35 | this be | 1,631 | -72.30 |
| that mean | 141 | 31.57 | we hope | 6 | -66.14 |
| labour need | 17 | 25.10 | we see | 4 | -64.99 |
| part be | 77 | 22.35 | we be | 57 | -49.95 |
| no one know | 27 | 20.50 | that-CL be | 30 | -40.45 |
| he tell | 45 | 20.36 | we learn | 6 | -40.00 |
| yesterday be | 36 | 19.45 | we hear | 3 | -33.34 |
| labour have | 32 | 19.07 | there be | 2,999 | -31.30 |

Our key colligation list tells us that *that* and *things* were more frequently used here as subjects together with copular verbs. Importantly, the list contains several evaluative colligations: most of *thing be* were followed by evaluative adjectives (e.g., *a thing is clear*.), which suggests that the previous contexts captured by a vague word, *thing* were evaluated. The same goes for a similar colligation, *that mean*, which paraphrases the previous context, including a writer's evaluation. A similar observation is made about *yesterday be*, which I checked manually, to find it still significant; this too has a definitive function of looking back on the previous day.

(3) In many regards, ***yesterday was*** just another bad day for British transport. (*The Guardian*, March 12, 2010, "High-speed rail: All aboard!", boldfaced by the author)

Thus, *The Guardian* tends to use summative phrases as subjects and verbs, as seen in *things be* and *that mean,* all of which are related to evaluations.

*The Daily Telegraph* also revealed an asymmetrical image but not the same one as *The Times* and *The Guardian* with their absolutely non-human preference (see Table 5.4):

Table 5.4: The top key colligations of *The Daily Telegraph*

| Positive key colligations | RF | LL | Negative key colligations | RF | LL |
|---|---|---|---|---|---|
| we be | 195 | 123.41 | there be | 1,988 | -142.67 |
| we have | 198 | 109.89 | this be | 1,258 | -48.16 |
| we suspect | 48 | 89.59 | question be | 67 | -47.67 |
| we hope | 84 | 68.57 | it be | 3,413 | -46.20 |
| we trust | 39 | 62.41 | issue be | 20 | -39.40 |
| we say | 34 | 53.84 | it would be | 243 | -37.05 |
| we argue | 35 | 51.04 | risk be | 7 | -28.78 |
| we believe | 41 | 40.51 | result be | 46 | -28.11 |
| we want | 30 | 36.79 | first be | 54 | -23.33 |
| we report | 32 | 35.96 | it will be | 86 | -22.81 |

The dominance of first person plurals at the top 10 means that *The Daily Telegraph* hit a strikingly higher frequency of this type of pronoun than the three other newspapers. In contrast, impersonal constructions (e.g., *there be*, *it be*, *it would be*, *it will be*) and non-human colligations (e.g., *questions be*, *issue be*, *risk be*, *result be*, *first be*) were downplayed, or back-grounded. It should be noted that our key colligation analysis enables us to analyse a variety of *we*-colligations at the significance level, leading to the discovery of exploitation of first person plurals by *The Daily Telegraph*.

In the continuum of personality and impersonality, *The Independent* is located in the middle, as Table 5.5 shows:

Table 5.5: The top key colligations of *The Independent*

| Positive key colligations | RF | LL | Negative key colligations | RF | LL |
|---|---|---|---|---|---|
| it be | 5,685 | 194.31 | that be | 623 | -42.13 |
| there be | 3,887 | 186.69 | aim be | 21 | -23.30 |
| this be | 2,429 | 181.20 | he tell | 8 | -17.93 |
| we tell | 99 | 136.63 | notion be | 4 | -15.53 |
| we need | 172 | 136.21 | he think | 4 | -15.08 |
| we must hope | 54 | 120.44 | budget be | 1 | -12.15 |
| fact be | 169 | 111.03 | challenge be | 29 | -11.88 |
| we see | 102 | 110.24 | case be | 37 | -11.57 |
| we have | 231 | 109.26 | issue be | 50 | -11.56 |
| truth be | 194 | 71.00 | he believe | 12 | -11.35 |

Since the key colligation analysis indicates that *it* and *there* were used in impersonal constructions without seeing their collocates, we have a clear picture that the top three colligations are impersonal but that five *we*-colligations are ranked in the top 10, while those including the third person singulars and particular abstract nouns turned up in the top 10 negative key colligations.

We have seen the top key colligations in each newspaper corpus, but the key colligation analysis enables us to reveal them in their true colours, using the grammar information such as the voice, modal auxiliaries. As one of the top *we*-colligations indicates, both *The Daily Telegraph* and *The Independent* made most of first person plurals, but in a different way, particularly as regards modality. *The Independent* frequently combined this type of pronouns with modals such as *should* and *must* at a significant level (see Table 5.6).

Table 5.6: Comparing key colligations of *we should* and *we must* between *The Daily Telegraph* and *The Independent*

| | *The Daily Telegraph* | *The Independent* | LL | *p*-value |
|---|---|---|---|---|
| we should | 145 | 325 | -32.66 | $p < .0001$ |
| we must | 65 | 244 | -70.06 | $p < .0001$ |

The statistics lend supports to our analysis; *The Independent* hit 54 instances of *we must hope* in sharp contrast with only four occurrences in *The Daily Telegraph* ($p<.0001$). The qualitative analysis indicates that *The Daily Telegraph* employed *we*-colligations to state its views explicitly and frequently with manner adverbs in the last paragraph of an editorial text, while *The Independent* manipulated the modality explicitly, presenting its suggestions as a consequence of such unavoidable circumstances. Compare the two excerpts from these two presses:

(4)  ***We*** earnestly ***hope*** that the Government's review does not significantly relax the laws governing IVF. Instead, it should lower the legal time limit for abortion and abolish the HFEA, replacing it with a permanent consultative committee of moral philosophers, theologians and scientists reporting to Parliament.  (*The Daily Telegraph*, August 17, 2005, "Let us debate scrapping the HFEA", boldfaced by the author)
(5)  Barack Obama has a daunting list of problems in his in-tray. ***We must hope*** he realises a change of strategy on the Israeli-Palestinian conflict is a priority and that he can discourage his Israeli allies from using the kind of bloody but counterproductive tactics to which they are now resorting.  (*The Independent*, December 29, 2008, "The bombardment of Gaza will destroy lives, not Hamas", boldfaced by the author)

In (4) *The Independent* stated its view straightforwardly, appealing strongly to its readers, while in (5) *The Independent* was a little modest, considering probable counterarguments and insisted that its decision was circumstantial, dodging full commitments.

Given the impersonal constructions with the same verb, the different modality of *The Independent* becomes clear. *The Independent* also employed the impersonal construction, *it is to be hoped*, which conveys more or less the same proposition but quite different attitudes towards it from a colligation of *we hope*, as in (6):

(6)  The Tory leadership election is still very much open. The activists have a clear-cut choice. For the sake of the country, which desperately needs a strong opposition party, ***it is to be hoped*** they choose more wisely than last time.  (*The Independent*, November 5, 2005, "The Tories cannot afford to make another foolish choice", boldfaced by the author)

This text occurred in the last paragraph, where writers end an editorial text with their own view. The use of an impersonal construction here helps a writer to mention her/his hope, while giving a fair take to the statement about the election (i.e., detachment). *The Independent* seems to skillfully switch these different modes. Unlike such double uses of personality and

impersonality with modals by *The Independent*, *The Times* made most of the impersonal construction including modality as seen in Table 5.7:

Table 5.7: The frequency of *it is to be hoped* and *we hope* among the British editorial corpora

| | *The Times* | *The Guardian* | *The Daily Telegraph* | *The Independent* |
|---|---|---|---|---|
| it is to be hoped | 38 (1.73) | 10 (0.49) | 7 (0.45) | 47 (2.33) |
| we must hope | 1 (0.05) | 0 (0) | 2 (0.13) | 56 (2.77) |

At the other extreme, *The Daily Telegraph* preferred to use *we hope* without modals in the present tense (97.37% of all the occurrences of this colligation) in the last paragraph (68.42%). Moreover, this use is accompanied by attitudinal adverbs such as *earnestly*, *fervently*, *sincerely*, *urgently*, and *devotedly*, which places an emphasis on authority (see (4)). In fact, this medium frequently combined the first person plurals with verbs of saying, which only editorialists can use in this type of text (see Table 5.8):

Table 5.8: *We*-colligations in *The Daily Telegraph* and *The Independence*

| Colligations | *The Daily Telegraph* | *The Independent* | LL | *p*-value |
|---|---|---|---|---|
| we report | 32 (2.04) | 14 (0.70) | 12.54 | $p < .001$ |
| we say | 29 (1.85) | 10 (0.50) | 15.11 | $p < .001$ |
| we argue | 34 (2.17) | 10 (0.50) | 20.63 | $p < .0001$ |
| we urge | 17 (1.08) | 2 (0.10) | 17.66 | $p < .0001$ |
| we mean | 14 (0.89) | 16 (0.79) | 0.11 | n.s. |
| we are told | 16 (1.02) | 78 (3.90) | -30.33 | $p < .0001$ |

Contextualising the colligations tells us that the writers stressed their presence; they have long been engaged in reporting, which is supported by the present tense or the present perfect with time adverbs.

(7) a. ***We report*** today a statistic that should make the Government blush with shame. (*The Daily Telegraph*, March 22, 2010, "A shocking statistic that sums up Britain's plight", boldfaced by the author)
b. ***We have argued again and again*** that the events of the past few months demand an election this autumn. (*The Daily Telegraph*, July 25, 2009, "Voters are ready to turn off Labour's life support", boldfaced by the author)
c. ***We say "unbelievably"*** because that case illustrates, more graphically than any other, the utter incompetence of the government bureaucracies that now want additional powers. </p> (*The Daily Telegraph*, June 26, 2006, "Bureaucracy can't bring up children", boldfaced by the author)

In the last example, the editorialists called for caution towards their wordings, which is often vital to journalism. Another difference is the passive in which *The Independent* employed *we*-colligations. The passive use means elusive, unclear attitudes towards commitment or textual responsibility.

(8) ***We are constantly told*** that troop withdrawals will begin when Iraq's security forces prove able to take on more responsibility themselves.
(*The Independent*, October 23, 2006, "A moment of clarity amid the bloodshed and carnage; IRAQ", boldfaced by the author)

Such a time adverb as *constantly* conveys the message that the writer and readers are a little sick of the constant announcement of something by the government.

The passive in *The Daily Telegraph* was limited to colligations of first person plurals and psych predicates, where writers expressed themselves. *The Independent* had a similar usage but with a modal auxiliary (e.g., *we should be pleased*); making the purposes of its passive seem to circumvent responsibility, represented by the agentless passive and the receptive use, mentioned above. For more, see Table 5.9:

Table 5.9: The key colligations with the passive

| *The Times* | *The Guardian* | *The Daily Telegraph* | *The Independent* |
|---|---|---|---|
| it can be claimed,<br>it can be assumed,<br>he should be prepared,<br>he be sentenced,<br>people be killed,<br>it should be acknowledged | it be called | vacuum be filled,<br>we be delighted,<br>we be promised,<br>we be pleased | we be told,<br>it should be noted,<br>it be estimated,<br>we should be pleased,<br>it could be argued,<br>he be justified,<br>it can be excluded,<br>it must be recognised,<br>we be faced,<br>question be raised |

As shown in Table 5.9, the top passive colligations tell us that *The Times* and *The Independent* also employed the agentless passive, which downplays writers' profile and makes writers' responsibility unclear. In fact, Table 5.10 shows that *The Times* revealed the greatest variety and the highest frequency among the four quality newspapers.

Table 5.10: The types and tokens of the passive colligations

| Newspapers | Types | LL | *p*-value | Tokens | LL | *p*-value |
|---|---|---|---|---|---|---|
| *The Times* | 8,734 (3967.13) | 163.35 | $p<.0001$ | 9,926 (4508.55) | 191.49 | $p<.0001$ |
| *The Guardian* | 6,528 (3,188.82) | -93.10 | $p<.0001$ | 7,188 (3511.21) | -172.96 | $p<.0001$ |
| *The Daily Telegraph* | 5,094 (3,247.68) | -44.99 | $p<.0001$ | 5,713 (3642.33) | -65.54 | $p<.0001$ |
| *The Independent* | 7,305 (3613.98) | 5.59 | $p<.05$ | 8,552 (4230.91) | 35.01 | $p<.0001$ |

Note: The number in ( ) represents the normed frequency.

The number of passive divides British quality newspapers into two: *The Times*/*The Independent* and *The Daily Telegraph*/*The Guardian*. *The Guardian* and *The Daily Telegraph* showed fewer uses of the passive than the other two. The underuse of the passive in *The Daily Telegraph* is possibly due to its personal, hortatory style, which is not true of *The Guardian*.

As we have seen, *The Guardian* exhibited a unique impersonal style. The key colligation analysis, which deals with lexical items in the massive data, uncovered the categorical analytical features of *The Guardian*. First, this medium was characterised by general reference, as seen in (9) and Table 5.11:

(9) <p> Yesterday, the chief police officers at their annual conference conceded that the new act would be "a welcome and civilised approach to drinking" for the vast majority of people, but warned that relaxed pub hours could lead to a rise in violent crime. ***Research suggests*** drink is involved with 40% of domestic violence, 65% of murders and 75% of stabbings. (*The Guardian*, May 12, 2004, "Drinking hours: Bar counsel", boldfaced by the author)

Table 5.11 General reference in *The Guardian*

| | |
|---|---|
| $p<.001$ | research suggest, source say, poll find, poll put |
| $p<.01$ | analysis suggest, report document, poll say, evidence show, survey be, poll show |
| $p<.05$ | survey find, statistics be, report warn, survey suggest, report note |

In the above excerpt, the writer cited *research* to support the warning by the police with evidence, but the news source is not mentioned at all; it may come from the news reporting in the same newspaper or from beyond the text, which is called 'intertextuality' (for intertextuality, see Fairclough, 1992, Hunston, 2001 and Teubert, 2000). However, the word itself is so authoritative as to gain credibility from readers. Writers used this device to lend support to their argument, while anticipating possible counter-arguments or different ideas from their own which they could forestall. We call this 'general reference,' which is easy for knowledgeable writers to create, but hard to rebut, because there are no visible contenders.

Intriguingly, a similar analysis was carried out by Hunston (1993: 107) from her study of research articles in *Language in Society*. In academic papers, there were two ways of referring to other researchers; non-specific nouns of grouped specialism, such as *sociolinguists* and general nouns denoting academic fields, such as *studies,* referring metonymically to scientists. 'Internal sources' as *findings* in *The findings presented in Table 1 show …*, were also noticeable there. Hunston raises three points; other researchers and internal sources serve self-sustaining argument and an already-interpreted sort of impeccable truth and the process of knowledge construction. According to Hunston (1993: 107), the construction of knowledge is intertextually processed, because the writer's knowledge and claims are always based on other researchers' claims. For solid validity, more researchers as sources are necessary. This is true of arguments in editorials. It should be noted that in her above example, "findings" was not unspecific but most of our instances of general reference were non-specific, possibly because the editorials are not so strict with attribution as research papers are (cf. Bednarek, 2006: 53; Charles, 2006: 315; Tadros, 1993: 107).

*The Guardian* was also characterised by another type of reference, that is, the combination of quantified subjects and verbs of saying: what this paper calls 'anonymous voicing'.

(10) <p> In Russia ***no one can*** quite ***understand*** how the national image abroad has fallen so far and so rapidly. ***Some say*** it is all a plot hatched by oligarchs exiled in London. ***Others say*** that the west was more than happy to engage with Russia when it was weak under President Boris Yeltsin, but cannot face competition from a strong Russia under Mr Putin. (*The Guardian,* February 9, 2007, "Russia: Turning colder", boldfaced by the author)

A pair of *some say* and *others say* is extremely common in persuasive prose for well-balanced argumentation to remove a one-sided view and show a certain objectivity. However, it should be noted that this cliché is actually groundless and suggests the control of argumentation by *The Guardian*. This supports the view of Bhatia (1993) that in sharp contrast with news reporting editorials are biased in favour of one's newspaper; however, it is also true of other argumentative genres. Miki (2010) mentions that LOCNESS, a collection of argumentative essays by American and British students had a higher frequency of a quantifier, *many* than other genres of written prose (i.e., FROWN and FLOB) to create a particular semantic prosody; *many*

*people* was used for a temporary generalisation, which was denied later. The same sematic prosody of *many* was found in my TOEFL model essay corpus (Miki, 2009). Hunston (1993) also indicated the argument-construction with such vague expressions in her analysis of radio discussion in 'Any Questions,' where all four of the monologues answered a single question; in radio discussion, there were 'vague words' such as *everybody* and *nobody,* to which some disputable opinions were attributed. See the citation from Hunston:

> In other words, the political world is observed and judged by imaginary ordinary people, possessing superior knowledge and experience. Given that the audience to the programme consists of such 'ordinary people', a motivation for this is not far to seek, but it is worth noting that this is one of the ways in which speakers construct, and acknowledge the face of, their audience. (Hunston, 1993: 106)

It is no coincidence that newspaper editorials show the same tact. Since newspaper readership is composed of ordinary people, editorialists develop their argument with high awareness of popular views. That is, they aim at developing an argument which can be enhanced or brought to balance by referring to general knowledge and other views. Writers of persuasive prose take advantage of this to incorporate unsourced information into their argument.

Another categorical feature of *The Guardian* is numbers. This medium had a journalistic intake with numerals (i.e., Arabic numbers). It scored a much higher frequency of numerals than any other newspaper at a significant level, while *The Independent* hit the bottom (see Table 5.12):

Table 5.12: The frequency of numerals and numeral subjects in the corpora of British newspaper editorials

| | Numerals | | | Numeral subjects | | |
|---|---|---|---|---|---|---|
| | R.F. (N.F.) | LL | *p*-value | R.F. (N.F.) | LL | *p*-value |
| *The Times* | 19,310 (8,770.92) | 68.49 | $p < .0001$ | 188 (85.39) | -43.22 | n.s. |
| *The Guardian* | 20,590 (10,057.86) | 951.69 | $p < .0001$ | 306 (149.48) | 163.55 | $p < .0001$ |
| *The Daily Telegraph* | 13,161 (8,390.80) | 0.56 | n.s. | 94 (59.93) | -21.64 | $p < .0001$ |
| *The Independent* | 12,309 (6,089.60) | -1782.53 | $p < .0001$ | 151 (74.70) | -15.48 | $p < .0001$ |

Including numerals followed by *%* and *percentage*.

Interestingly, *The Times* had a significant number of numerals overall but not as a subject, a feature which is true of *The Daily Telegraph* also. This indicates that the subject nouns were highly selective. The use of numerals as a subject reminds us of academic writing. Compare the following pairs.

(11) a. Over ***80***% of the people interviewed by Oxfam in the Eastern Congo ***said*** that security was worse now compared to a year ago. (*The Guardian,* July 20, 2009, "Congo: Perilous peace", boldfaced by the author)

b. Figure 2 shows the Fe spectra of the Al substantial alloys taken at 77K (Gosden, 1993: 66)

A number such as a percentage makes it authentic and factual. Gosden (1993) indicates that this kind of subject is frequent in the discussion sections of academic papers. To put it another way,

it is possible that this kind of subject is often part of the explanation of visuals in discussion sections. Similarly, numeral subjects in editorials are typically statistics from surveys or questionnaires but without visuals. This genre provides readers with textual information rather than visuals, resulting in the different specificity between research papers and editorials; however it is still a piece of hard evidence in argumentation.

Considering the functional role of the subject in texts, numeral subjects are so foregrounded that they can have a huge impact on the interaction of readers and writers, and are also established as pieces of hard evidence. Thus, in terms of numeral subjects, writing in *The Guardian* is similar to academic writing maybe in the field of hard science but is different from it with respect to anonymous voicing.

The key colligation analysis found that non-human lexical subjects were exploited in all the newspaper editorials but *The Daily Telegraph.* In particular, there are two notable types of lexical noun in the top 10 colligations: divulging and labelling. Divulging denotes information disclosure; writers make explicit statements against readers' expectations or inference from a previous context.

(12)  It is perfectly true, as official figures last week demonstrated, that unemployment is once again on the rise in Britain. And the numbers of those seeking work are only likely to increase as the shock of the debt meltdown works its way though (sic.) our economy. But the ***truth is*** that Britain's flexible immigration policies of recent years are likely to help matters, rather than make them worse. (*The Independent*, October 20, 2008, "A proposal that would make a bad immigration policy worse", boldfaced by the author)
(13) There is now a $1.2bn security programme to foil Olympic bombs of every kind, dirty, clean, human, inhuman. And in case all that fails, the International Olympic Committee has bought 95m of cover to insure themselves against cancellation. …. One ***thing is*** already clear. This year's games will be the most militarised yet.  (*The Guardian*, October 20, 2008, "A proposal that would make a bad immigration policy worse", boldfaced by the author)

The frequent combination of divulging consists of a general noun (e.g., *truth*, *thing*) and a copular verb, whose combination is neutral but signals the propositional contents to come after them. The divulging colligations are characteristic of *The Independent* and *The Guardian*.

In contrast, labelling is more cohesive and more evaluative, though the difference between labelling nouns and divulging is sometimes subtle; the previous studies of labelling nouns insisted on summative and evaluative functions (Francis, 1994). See the labeling colligations such as *the first issue is* and *the other issue is* in (14):

(14) Such services, however, are as controversial as they are innovative. And they raise ethical, legal and bureaucratic questions that must be resolved if such genetic advances are to benefit, rather than terrify, society. ***The first issue is*** regulation. ….
***The other issue is*** the danger of insurance companies misusing this service to discriminate against those whose profile suggests the early onset of catastrophic disease. (*The Times*, March 1, 2008, "The Inner Self", boldfaced by the author)

*Issue* in these labelling colligations is linked to the precedent words such as *controversial* and *questions*. To put it another way, the labeling word summarises the previous text and expresses evaluation; *issue* implies a complicated matter. More importantly, it is no coincidence that the significant number of labelling nouns occurred in the subject position, as Yamazaki (2008) mentions:

> … a typical position of a label within a clause structure: the label occupies the subject position, followed by a verb phrase. Labels can basically occur anywhere in a sentence, but when presenting given information for a cohesive purpose, ***they tend to occur within the first part of a sentence or the theme position*** (Halliday 1994: 37–48). This position is so typical of labels that it can be used as one of the means for searching for labels in texts. (Yamazaki, 2008: 80, boldfaced by the author)

Although the occurrences of a label in the subject can be attributed to its cohesive nature, the actual use was limited to particular papers; at the top key colligations, *The Times* and *The Independent* overused labelling in comparison with *The Guardian* and *The Daily Telegraph*. It can be interpreted that the former papers tend to develop the logical argument over the length of texts for implicit evaluation. In contrast, the latter were likely to express their views in a straight, direct way.

We have demonstrated the key colligation analysis from four British quality newspapers with a focus on sequences of subjects, verbs and auxiliaries, if any, distinguishing active/passive voices. Our unique analysis, followed by the qualitative analysis, unveils the categorical features as summarised in Table 5.13:

Table 5.13: Summary of the key colligations analysis of the British newspaper editorials

| Corpora | Features of key colligations (subject, verbs, auxiliaries, the voice) |
|---|---|
| *The Times* | lexical human subject, *he*, impersonal constructions (i.e., *there be*), the passive (incl. agentless passive), labelling subjects |
| *The Guardian* | general reference, anonymous voicing, numeral subjects, divulging |
| *The Daily Telegraph* | *we* with verbs of saying, attitudinal adverbs in the present tense, the passive (incl. *we*) |
| *The Independent* | *we* with modal auxiliaries, the passive (incl. *we* and the agentless passive), impersonal constructions (i.e., *there be* and *it would be*), divulging |

The findings indicate that bundles of linguistic features can indicate styles (cf. Biber, 1998): authority represented by *we*; impersonality which took the form of *there* constructions and the agentless passive, intertextuality—general reference and anonymous voicing, for instance. These categorical features characterise newspaper editorials but also reveal the professional manipulation of styles within the same mass media.

## 6. Conclusion

This methodology-based paper has first introduced key colligation analysis to characterise the editorials of the selected British newspapers in the light of subjects and predicates. The way of identifying characteristics is similar in several ways to keyword analysis (Scott, 2000, 2001, 2010); it identifies the specialised words by comparing two wordlists of a target corpus and a reference corpus on the basis of the log-likelihood ratios. Second, the analysis compares the positive (statistically more frequent) and negative (statistically less frequent) words. However our key colligation analysis is differentiated from the keyword analysis in dealing with phrases with particular grammatical relations. It should be noted that key cluster analysis (Baker, 2006) can cover a sequence of words, but it does so at random without targeting any particular grammatical relations.

The results of the key colligation analysis showed how effectively and efficiently this method captured the significant combinations of subjects and predicates, which had an impact on the style or the interpersonal relations between writers and readers. In fact, our findings are

keys for differentiating the British quality newspapers. *The Times* took advantage of third singular nouns such as countries, *he*, which suggests celebrities and politicians as well as revealing impersonality by expletives and labelling, which implies a highly developed textual structure. *The Daily Telegraph* exploited first person plurals sometimes to exclude readers (e.g., *we hope* …) for evaluation or at other times to include them (e.g., *we are constantly told* …) for involvement. *The Independent* had double faces; it foregrounded or back-grounded the writers, switching between the first person plurals and the agentless passives for detachment. *The Guardian* is adept at developing argumentation. In what we call anonymous voicing, writers invent unknown, non-specific, invisible people to state their views or the opposite and provide several contrastive but well-balanced views to readers. Also, this research, based on massive amounts of persuasive prose, gives validity to the result of the discourse analysis of argumentative development in academic papers and radio programmes (Hunston, 1993) and the divisions of Hortatory Exposition and Analytical Exposition by Martin (1985a).

Generally, quality newspapers are used as a precious resource of investigation of English usage; some researchers recommend learners to see editorials as models of argumentative prose (Connor, 1996). As we have seen, there are stylistic variations of argumentation even in the same genre of editorials. In spite of this, the BNC, one of the most popular databases of English texts, contains editorials only from *The Independent* according to the BNC index (http://tiny.cc/davidlee00, see also David, 2001). There is a possibility that the users of the BNC judge the characteristics of newspaper editorials, based on a single newspaper. The present study, as an implication, has issued a warning in this respect. Another implication of this research is that the same approach can be applied to the investigation of verbs of attribution and the subjects in other fields, such as academic writing. Several studies of this have been done but the range is limited due to the methodology; researchers first set target verbs or nominative pronouns and investigate the actual uses in contexts or carry out discourse analysis of the relatively small number of data. In contrast, our key colligation analysis is genuinely data-driven, quantitative investigation, followed by discourse analysis.

**Acknowledgement**
This paper was based on Miki (2011), my PhD thesis in Osaka University.

**References:**

Baker, P. (2004). Query keywords, questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, 32(4), 346–359.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Baynham, M. (1999). Double-voicing and the scholarly 'I': On incorporating the words of others in academic discourse. *Text*, 19(4), pp. 485–504.

Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus*. London: Continuum.

Bhatia, V. K. (1993). *Analyzing genre: Language use in professional settings*. London: Longman.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

Cadman, K. (1997). Thesis writing for international students: A question of identity? *English for Specific Purposes*, 16(1), 3-14.

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of study of theses in two disciplines. *English for Specific Purposes*, 25, 310–331.

Connexor Oy (2008). *Connexor Machinese Syntax for English* (Ver. 3.9.3.4) [Computer software]. Helsinki: Connexor Oy.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing*. Cambridge, England: Cambridge University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.

Firth, J. R. (1957 [1951]). Modes of meaning. In F. Palmer (Ed.), *Selected papers of J. R. Firth 1934-1951* (pp. 190-215). London: Longman.

Fortanet, I. (2004). The use of 'we' in university lectures: Reference and function. *English for Specific Purposes*, 23, 45–66.

Francis, G. (1989). Thematic selection and distribution in written discourse. *Word*. 40(1-2), 201–221.

Francis, G. (1994). Labelling discourse: An aspect of nominal-group lexical cohesion. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 81–101). Abingdon, Oxford, England: Routledge.

Fries, H. P. (1981). On the status of theme in English: Arguments from discourse. *Forum Linguisticum*. 6, 1–38.

Fries, H. P. (1983). On the status of theme in English: Arguments from discourse. In János S. Petöfi & Emel Sözer (Eds.), *Micro and macro connexing of texts* (pp. 116–152). Hamburg, Germany: Helmut Buske Verlag.

Fries, P. H. (1994). On theme, rheme and discourse goals. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 227–249). Abingdon, Oxford, England: Routledge.

Gosden, H. (1993). Discourse functions of subject in scientific research articles. *Applied Linguistics*, 14(1), 56–75.

Goto, K. (2005). Eigo collocation no chushutsu ni kansuru Machinese Syntax to Link Grammar Parser no hikaku kenkyu. [Comparison of Machinese Syntax and Link Grammar Parser in retrieving collocations]. *Osaka Studies in Corpus Linguistics 2004-2005* (pp. 32–42). Graduate School of Language and Culture, Osaka University.

Goto, K. (2006). Gakujyutsu eigo Corpus niokeru SV hyogen chushutsu no kokoromi: Corpus based EAP learning, (retrieval of SV expressions from the academic corpora: Towards corpus-based learning). *Osaka Studies in Corpus Linguistics 2005-2006* (pp. 3–12). Graduate School of Language and Culture, Osaka University.

Goto, K. (2008). Denshi corpus kaiseki ni motozuku eigo collocation kenkyu: corpus gengogaku ga motarasu kokatekina collocation gakushu no kanosei (A study of English collocations based on corpus parsing: The possibility of effective learning by corpus linguistics). PhD thesis at Osaka University.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English. *Journal of Linguistics*, 3, 199-244.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Hawes, T. & Thomas, S. (1996), Rhetorical uses of theme in newspaper editorials. *World Englishes*, 15(2), 159–170.

Hindle, D. (1994). A parser for text corpora. In A. Zampolli (Eds.), *Computational approaches to the lexicon*, (pp. 103–151). NY: Oxford University Press.

Hoey, M. (2005). *A lexical priming: A new theory of words and language*. Abingdon, Oxford, England: Routledge.

Hunston, S. (1993). Projecting a sub-culture: The construction of shared worlds by projecting clauses in two registers. In D. Graddol, L. Thompson & M. Byram (Eds.), *Language and culture* (pp. 98-112). Clevedon: BAAL/Multiingual Matters.

Hunston, S. (2000). Evaluation and the planes of discourse: status and value in persuasive texts. In S. Hunston & G. Thompson (Eds.), *Evaluation in text*, (pp. 176–207). Oxford, England: Oxford University Press.

Hunston, S. (2001). Colligation, lexis, pattern, and text. In M. Scott & G. Thompson. (Eds.), *Patterns of text: In honour of Michael Hoey* (pp. 13–33). Amsterdam: John Benjamins.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes*, 20, 207–226.

Hyland, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics*, 34. 1091–1112.

Ishikawa, S. (2008). *Eigo corpus to gengo kyoiku: data toshiteno text*. [English corpora and language education: texts as data]. Tokyo: Taishukan.

Ivanič, R. (1998). *Writing and identity: The discoursal construction of identity in academic writing*. Amsterdam: John Benjamins.

Kuo, Chin-Hua. (1999). The use of personal pronouns: Role relationships in scientific journal articles. *English for Specific Purposes*. 18(2), 121–138.

Íñigo-Mora, I. (2004). On the use of the personal pronoun *we* in communities. *Journal of Language and Politics*, 3(1), 27–52.

Lee, D. Y. W. (2001). Genres, registers text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3), 37-72.

Martin, J. R. (1985a). *Factual writing: Exploring and challenging social reality*. Oxford, England: Oxford University Press.

Martin, J. R. (1985b). Exposition: Literary criticism. In J. R. Martin (Ed). *Factual writing: Exploring and challenging social reality* (pp. 83–99). Oxford, England: Oxford University Press.

McCrostie, J. (2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal*, 32, 97–114.

Miki, N. (2009). *The influence of choice of reference corpora on the results of keyword analysis*. Unpublished M.Phil thesis. The University of Birmingham, the UK.

Miki, N. (2010a). Numerals and quantifiers in argumentative writing. *Journal of Language and Culture*, 19, 53–68. Society for the Study of Language and Culture, Osaka University.

Miki, N. (2010b). The SV expressions as a rhetorical device from the British broadsheet editorials. In A. Harris & A. Brandt (Eds.), *Language, Learning and Context: Proceedings of the BAAL Annual Conference 2009* (pp. 99–103), 42nd Annual meeting of the British Association for Applied Linguistics, 3-5 September 2009, Scitsiugnil Press. [CD-ROM]

Miki, N. (2011). A corpus analysis of argumentative writing: Authority and intertextuality in the British newspaper editorials. Ph.D. thesis. Osaka University, Japan.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh, Scotland: Edinburgh University Press.

Oktar, L. (2001). The ideological organization of representational processes in the presentation of us and them. *Discourse & Society*, 12(3), 313–346.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 106–118). London: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In Granger, S. (Ed.), *Learner English on computer* (pp. 41–51). London: Longman.

Scott, M. (2000). Focusing on the text and its key words. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 103–121). Frankfurt, Germany: Peter.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 47–67). Amsterdam: John Benjamins.

Scott, M. (2010). *WordSmith Tools* (Version 5.0) [Computer software]. Oxford, England: Oxford University Press.

Scott, M. & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Tadros, A. (1993). The pragmatics of text averral and attribution in academic texts. In M. Hoey (Ed.), *Data, description, discourse: Papers on the English language in honour of John McH. Sinclair* (pp. 98–114). London: HarperCollins.

Tapanainen, P. (1999). *Parsing in two frameworks: finite-state and functional dependency grammar*. Unpublished PhD thesis. University of Helsinki, Finland.

Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogical account of the language of project proposals. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75–90). Frankfurt, Germany: Peter.

Teubert, W. (2000). A province of a federal superstate, ruled by an unelected bureaucracy: Keywords of the Eurosceptic discourse in Britain. In A. Musolff, C. Good, P. Points & R. Wittlinger (Eds.), *Attitudes towards Europe: Language in the unification process* (pp. 45-86). Aldershot, Hampshire, England: Ashgate.

Ventola, E. (1994). Finnish writers' academic English problems with reference and theme. *Functions of Language*, 1(2), 261–293.

Yamazaki, N. (2008). Collocations and colligations associated with discourse functions of unspecific anaphoric nouns. *International Journal of Corpus Linguistics*, 13(1), 75–98.

Westin, I. (2002). *Language change in English newspaper editorials*. Amsterdam: Rodopi.