Use of signalling nouns across L1 and L2 writer corpora*

John Flowerdew
University of Leeds / City University of Hong Kong

A signalling noun is the use of an abstract noun the full meaning (realisation) of which can only be made specific by reference to its linguistic context. Examples of nouns which can function as signalling nouns are attitude, consequence, difficulty, effect, process, reason and result. The realisation of a signalling noun may occur across clauses or within the clause. Using the ICLE Locness (L1 writers) corpus as a reference corpus and a learner English corpus written by Cantonese-speaking learners of English, this paper presents a comparison of the use of signalling nouns by these two L1 and L2 writer groups. Use of different functions, realisation patterns and selection and range of individual nouns are compared. In this way a profile is built up of the use of signalling nouns by the two target groups and areas for pedagogic intervention are identified.

Keywords: signalling noun, abstract noun, learner corpus, cohesion, lexical cohesion

1. Introduction

Signalling nouns (SNs) are abstract nouns which have particular cohesive properties. A SN is potentially any abstract noun the full meaning of which can only be made specific by reference to its linguistic context. The term ‘signalling’ is meant to indicate this cohesive property. Examples of nouns with a potential for functioning as SNs include the following: attitude, consequence, difficulty, effect, process, reason, result. The signalling function of these nouns can operate across clauses — either cataphorically or anaphorically — or within the clause. Examples (1) and (2) illustrate the across-clause function with the noun fact, example (1) signalling cataphorically and example (2) signalling anaphorically:

(1) This theory leaves a number of facts unexplained. For example, starch is absent from the guard cells of certain plants; some guard cells lack...
chloroplasts but still open and close; and the stomatal movements of some plants may not necessarily be related to the time of day.

(2) Electricity is used to drive the motor of an electric train, but inevitably some of the energy is lost as heat. This unavoidable fact is of great importance in biology.

The relationship between a SN and its realisation in these examples is like the relationship of a pronoun to its antecedent; a SN signals that a lexical realisation may be found in its cotext in a similar way to how a pronoun signals that a referent may be found in its preceding cotext.

Examples (3)–(5) show the SNs chance, realisation, and role operating within the clause:

(3) a relatively low chance of proceeding to AIDS.

(4) the realisation that it was possible to simulate a prebiotic milieu in the laboratory ushered in a new era in origin-of-life studies.

(5) Their role is to carry out the depolarizing phase of an action potential.

In example (3) the realisation takes the form of a post-modifying of clause. In example (4) the realisation is as a post-modifying that complement clause. In example (5) the realisation is in the form of a non-finite clause introduced by the preposition to (see Section 2 below for other possible patterns).

A number of authors have discussed various aspects of what have been referred to by Flowerdew (2002, 2003a, 2003b, 2006) as SNs. Winter (1977) talks about “type 3 vocabulary”, emphasising the anaphoric properties of such nouns. Tadros (1985) discusses their predictive function. Francis (1986) focuses more on the anaphoric function of what she calls “a-nouns”, although she also considers the predictive function. Ivanic (1991) was probably the first to study the in-clause as well as the across-clause function, using the term ‘carrier noun’ (see also Winter 1992). Finally, Schmid (2000) is the most comprehensive treatment to date of the forms and functions of these nouns, both within the clause and (less so) across clauses, using the term ‘shell nouns’. Earlier, Halliday & Hasan (1976) included these nouns in their broader category of ‘general nouns’ (see Mahlberg 2005 for a more recent treatment). The key characteristic of all of these differently labelled categories is that there is an equative relation that relates the noun to its realisation in the cotext.

Some linguists have produced lists of some of the various categories of abstract nouns which are referred to in this study as SNs (e.g. Francis et al. 1998, Schmid 2000, Winter 1977), although these lists use different criteria for inclusion and do not agree with each other. These lists of SNs cannot be said to be finite, because nominalisation and lexical metaphor allow for the creation of new SNs.
Focussing on the in-clause function, Schmid (2000) takes a structural approach, basing his list on computerized corpus-based searches of the grammatical patterns in which the nouns typically occur. While Schmid provides a tremendous amount of data and examples of shell nouns, the computer searches are nevertheless limited and are not able to find all instances of any pattern. Neither do they take account of the across-clause function. Schmid’s method, thus, does not allow for the possibility of counting the overall frequency of shell nouns.

The contrasting approaches to this class of abstract noun that is here referred to as SN and the problem of retrieving all instances of those nouns, even if those nouns that belong to this class of nouns are specified in advance, means that any automated, replicable quantitative approach is problematic. Consequently the approach adopted here, working with a relatively small corpus, is to conduct the analysis by hand (assisted by automated corpus techniques). This is time-consuming, but nevertheless the most effective approach available (see Section 3 below).

Turning now to the literature on learner corpora, work with such corpora has been going on for well over a decade now (Granger 1998, 2002, 2004; Mukherjee, 2006). One of the goals of this research is to contrast various levels and varieties of L2 competence with an L1 model and to map sequences of development working towards the latter (Cobb 2003:395). Working within this framework, for example, Lorenz (1999) identified four categories of error by which learners may deviate from an L1 model: overuse, underuse, misuse and learner idiosyncratic forms. Nesselhauf’s (2005) analysis of collocations in learner English, which identified gaps between learner and L1-speaker usage, is another example of this type of research.

Also working within this framework, but specifically focussing on SNs, in an earlier paper, Flowerdew (2006) studied the use of SNs in a learner corpus of argumentative essays, showing how learner usage of these nouns correlated with grades awarded by the learners’ teachers, both in terms of frequency and accuracy of use, and thereby confirming the intuitive idea that SNs contribute to the coherence of a text. The present paper extends that study and compares the use of SNs in a learner corpus (Cantonese-speaking learners of English) with that of a comparable L1 writer corpus. It compares the learner usage of SNs with that of native speakers in terms of overall frequency, frequency of different functions (across-clause, cataphoric and anaphoric, and in-clause), selection of specific SNs, range of SNs selected for use, and frequency of different in-clause realisation patterns. In this way a profile is presented of the use of SNs by the two target groups, the aim being to show where pedagogic intervention might help in bringing the learners nearer to an L1 model, which, it is assumed, is likely to be more coherent.

SN usage can be viewed as a form of nominalisation (many SNs are in fact deverbal nouns), or ideational grammatical metaphor (Halliday & Matthiessen...
2004), and as such can be associated with advanced forms of literacy. Halliday & Matthiessen (2004:636) describe grammatical metaphor as follows:

Ideational [grammatical] metaphor is learned later by children and is not part of the grammar of ordinary, spontaneous conversation that children meet in the home and neighbourhood; rather, it is associated with the discourses of education and science, bureaucracy and the law. Children are likely to meet the ideational type of metaphor when they reach the upper levels of primary school; but its full force will only appear when they begin to grapple with the specialized discourses of subject-based secondary education.

The simpler non-nominalised forms referred to in this quotation are called ‘congruent realisations’ by Halliday and Matthiessen, while the nominalised ones are referred to as ‘grammatical metaphor’. Using a SN, belief, Halliday & Matthiessen (2004:639) give the following pair of examples to illustrate the congruent and metaphorical relationship.

(6) Congruent: People most strongly believe that there is no …

(7) Metaphorical: The strongest belief of all is that there is no …

Given the late acquisition of ideational grammatical metaphor among L1 writers, it can be hypothesised that SN usage is also likely to be problematic for L2 users of English who come to the more literate forms of English after the more conversational ones. This hypothesis is borne out by Flowerdew (2006), where difficulties of Cantonese-speaking learners of English with SNs were identified and where it was demonstrated that the higher levels of proficiency were identified with higher levels of frequency and accuracy in their use. It can be hypothesised therefore that, as compared to their L1 writer peers, Cantonese-speaking learners of English may well use SNs less frequently, use them in less complex ways, and employ a different and narrower range of SN types (i.e. different individual items).

2. Research Questions

The following research questions are designed to test the hypotheses outlined above with regard to SNs in the two corpora:

1. What are the relative overall frequencies?

Here the total numbers of SNs in each of the two corpora are compared, it being anticipated that the L2 corpus will contain less SNs than the L1 corpus.

2. What are the relative frequencies of the different functions?
Here, a distinction is made between two functions: across-clause, which can be either cataphoric or anaphoric (see examples 1 and 2) and in-clause functions (see examples 3–5). Thus, frequencies for the following categories are sought:

a. total across-clause
b. across-clause anaphoric
c. across-clause cataphoric
d. in-clause

Whether learners prefer more or less of any of these patterns as compared to the L1 users is an open question.

3. What differences are there in terms of the selection of specific SN types?

Here, the concern is with the difference in the relative frequency of the use of different individual nouns (types). Which are the SNs most commonly selected by the L1 writers and the Cantonese-speaking learners respectively? This comparison can be made by rank-ordering the SNs in each corpus in terms of frequency. This rank-ordering might show, for example, that certain SNs are used relatively more frequently in one corpus than the other. If the rank ordering of the use of individual nouns is very different, then learners are diverging from L1 usage. If, on the other hand, the rank ordering is very similar, then learners would be close to L1 usage.

4. What differences are there in the range of different SN types?

Here, the concern is with how many different individual SN types (as opposed to tokens) are used by the L1 and L2 writers. It is hypothesised that the L1 writers will use a greater variety of different individual SN types than L2 writers, indicative of a wider range of individual SN types at their disposal. This hypothesis is based on the fact that the L1 writers are likely to have a wider vocabulary range than the learners and that this difference will extend to the particular set of vocabulary that constitutes the SNs.

5. What are the relative frequencies of the different realisation patterns for the in-clause function?

Given that in the in-clause function a range of patterns can be used, the concern here is with these different in-clause patterns. The patterns are as follows (examples from the corpora used in this study):

(8) SN + of + np (e.g. a consideration of the sanctity and value of human life)

(9) SN + of + V-ing (e.g. the idea of banning of smoking)
Use of signalling nouns across L1 and L2 writer corpora

(10) SN + to + non-finite (e.g. a firm decision to heed or to disregard the pleas of opponents to corporal punishment)

(11) SN + that clause (e.g. the fact that the opponents make unfair generalizations)

(12) SN + be + identifier (e.g. an everyday example of this is Charles Barkley compared to Madonna)

(13) Identifier + be + SN (e.g. A great deal of people are arguing if students using credit cards is a good thing)

(14) SN + be + that clause (e.g. The next consideration is that students have many chances for academic travel overseas)

(15) SN + such as + np/V-ing (e.g. the drawback of using the card, such as getting into uncontrollable purchasing habits and have difficulties in paying off the balance of the card)

The purpose of this analysis is to find out if any patterns are more or less preferred by either of the two groups. To a degree, this may be dependent on the choice of individual nouns, but learners may favour nouns which take certain patterns, because they are more “congruent”, closer to the L1, or have been the focus of instruction. If the learners are using certain patterns more or less frequently than the L1 writers, then this would suggest a place for pedagogic intervention to approximate learner usage more to that of L1 writers.

To summarise, in seeking answers to these questions, a profile is presented of the use of SNs by the two target groups, the aim being to show where pedagogic intervention might help in bringing the learners nearer to an L1 model, which it is assumed, is more coherent and appropriate in terms of register.

3. Corpora and methodology

As already indicated, two comparable corpora of argumentative writing were used in this study. Specifically, these were a learner corpus of English written by Cantonese-speakers and an L1 corpus. Both corpora are or will be part of the International Corpus of Learner English (ICLE) (although the version based on English written by Cantonese-speaking learners of English varies slightly from the one in the ICLE, as it was obtained before being submitted to ICLE and essays in our earlier corpus scoring the highest and lowest grades were omitted [see Flowerdew 2006])

The learner sub-corpus consists of 111,558 words in total. It is made up of argumentative essays written by Cantonese-speaking learners of English who were predominantly in their first year at a Hong Kong university. There were altogether
217 essays which made up this sub-corpus, all between 252 and 834 words in length, with an average essay length of 514 words. The essays were written on nine different topics, all of which were argumentative in nature. All of the essays were written under test conditions and were timed. Reference materials were allowed to be used by the students as they wrote these essays, though any material in the essays that was directly quoted by students was not included in the corpus.

The L1 English sub-corpus, which was drawn from the larger ICLE corpus known as LOCNESS, consists of 110,537 words in total. The corpus is made up of argumentative essays written by American English L1 students who were predominantly between 17–22 years old. There were altogether 93 essays in the corpus, all between 449 and 2,927 words in length, with an average length of 1,188 words. The essays were written on a wide variety of topics \( (n = 61) \), but all were argumentative in structure. The essays differ from the learner corpus in that they were not written under test conditions, nor were they timed. Similar to the learner corpus, reference materials were allowed to be used by the students as they wrote these essays, though any material in the essays that was directly quoted by students was not included in the corpus.

In summary, the corpora were highly comparable in terms of genre, size and age of the contributors. On the other hand, there were differences in that, in contrast to the L1 essays, the learners’ essays were done under test conditions, were timed, and were considerably shorter.

Once the corpora had been created, all SNs occurring in the two corpora were identified by hand; SNs were identified by the author and a trained rater both using two procedures. For the in-clause patterns abstract nouns occurring in the possible structural patterns as listed under research question 5 were identified. For the across-clause patterns SNs were identified by locating abstract nouns for which the full meaning, or referent, was to be found in the surrounding co-text. Tagging took place for both of these two procedures. Following that, further tagging was done according to the various features which are relevant to the research questions set out above (e.g. across-clause anaphoric, across-clause cataphoric, and in-clause structural patterns). Tagging was done by introducing information by hand into the corpora. This was possible, although time-consuming, because of the relatively small size of the corpora. Thus codes such as ANAC (anaphoric SN across-clause) or CNAC (cataphoric across-clause), for example, were used to indicate anaphoric across-clause and cataphoric across-clause functions respectively. The various in-clause patterns were similarly tagged with abbreviations such as ICNP (in-clause SN + np), ICing (in-clause SN + V-ing), ICto (in-clause SN + to), etc.

After quite a lengthy training period, due to the complexity of the tagging system, inter-rater reliability was checked. In the few cases where there was disagreement, further training was conducted. The number of such cases was not large.
Use of signalling nouns across L1 and L2 writer corpora

43

because only two raters were involved and further training enabled discrepancies to be ironed out. Once tagging had been completed, it was a simple matter of using a word frequency counter to obtain frequency data on all of the patterns that had been tagged.

The frequencies of SNs were analysed in two ways. First, the general frequencies per thousand words or per hundred thousand words, were counted, with all of the texts in each corpus aggregated. Considering the fact that there were possible variations among the writers, the means of frequencies of SNs in individual texts in the two corpora were also compared; t-tests were carried out to indicate the significance of the differences.

4. Results

The results will be presented according to the five research questions. With regard to the first question, concerning relative overall frequencies of SNs in the two corpora, as shown in Table 1, the relative frequencies of SN use per thousand words indicate that the L1 writers of English use SNs approximately 17% more often than do the learners (42 vs. 36 per thousand). The mean frequency in individual texts by L1 English writers \((M = 50.4, SD = 29.32)\) is approximately 2.7 times that of learners \((M = 18.56, SD = 8.59)\). A t-test suggests a significant difference between the two groups, \(t = 14.7, p < .0001\).

<table>
<thead>
<tr>
<th>Frequency of SNs per thousand words</th>
<th>Aggregated mean frequencies</th>
<th>Individual text-based mean frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English L1 corpus</td>
<td>Learner corpus</td>
</tr>
<tr>
<td>Frequency of SNs per thousand words</td>
<td>42</td>
<td>36</td>
</tr>
</tbody>
</table>

With regard to research question 2, concerning the relative frequencies of the different functions, as shown in Table 2, these are reported under the four headings of total across-clause, across-clause anaphoric, across-clause cataphoric, and in-clause. Mean frequencies are provided per thousand words in each corpus, along with individual text-based mean frequencies.

a. total across-clause

Total across-clause realisations of SNs per thousand words are exactly the same for both L1 writers and the learners (24 per thousand words). However, when the
distribution of SNs in individual texts is taken into account, there is a significant
difference, with L1 English writers using 1.3 times more SNs ($M = 28.4$, $SD = 18.94$)
than the learners ($M = 12.4$, $SD = 6.20$), $t = 11.11$, $p < .0001$.

b. across-clause anaphoric
The L1 English writers’ corpus has a 31% higher frequency per thousand words
(17 vs. 13 per thousand) and a 194% greater mean by text ($M = 20.28$, $SD = 15.03$)
of anaphoric SN realisations in comparison to the learner corpus ($M = 6.89$, $SD = 4.48$). T-test results confirm a significant difference, $t = 11.73$, $p < .0001$.

c. across-clause cataphoric
The learners have a 57% higher frequency of cataphoric across-clause SN reali-
sations (11 per thousand) than that of the L1 English writers (7 per thousand). However, t-test results suggest an overall greater frequency of cataphoric across-
clause SN realisations in the L1 English writers’ corpus ($M = 8.67$, $SD = 5.39$) than
in the learner corpus ($M = 5.97$, $SD = 2.94$), $t = 5.58$, $p < .0001$, although the dif-
ference between the means (45% greater in L1 English writers) is not as significant as
in functions b and d.

d. in-clause
The L1 English writers use in-clause SNs 50% more frequently than the learners (18
vs. 12 per thousand). The mean by text in the L1 writer corpus is 2.4 times higher
($M = 22.08$, $SD = 14.05$) than that of the learner corpus ($M = 6.44$, $SD = 3.88$). T-test
results show a significant group effect in in-clause SN use, $t = 15.03$, $p < .0001$.

In summary for research question 2, t-test results confirm the intuitive hypothesis
that the learners use all of the above functions of SNs less frequently than the L1
English writers. The difference in terms of means of frequencies in individual texts
ranks from highest to lowest as follows: in-clause SN, anaphoric across-clause SN,
and cataphoric across-clause SN. The discrepancies between the overall frequencies per thousand words and the means of frequencies by individual texts suggest that there may be greater idiosyncratic variations in the learners’ use of SNs, particularly in cataphoric across-clause SN realisations, where the learners have an overall higher frequency, but the L1 English writers on average use them more frequently.

Concerning research question 3, the differences in terms of the selection of specific SNs, Table 3 shows the most commonly used SNs in each corpus ordered according to frequency.

Table 3. The most frequent SN types in each corpus

<table>
<thead>
<tr>
<th>No.</th>
<th>Native English Corpus</th>
<th>Cantonese L1 Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>argument</td>
<td>problem</td>
</tr>
<tr>
<td>2</td>
<td>problem</td>
<td>advantage</td>
</tr>
<tr>
<td>3</td>
<td>case</td>
<td>disadvantage</td>
</tr>
<tr>
<td>4</td>
<td>reason</td>
<td>method</td>
</tr>
<tr>
<td>5</td>
<td>claim</td>
<td>ban</td>
</tr>
<tr>
<td>6</td>
<td>way</td>
<td>way</td>
</tr>
<tr>
<td>7</td>
<td>act</td>
<td>result</td>
</tr>
<tr>
<td>8</td>
<td>example</td>
<td>example</td>
</tr>
<tr>
<td>9</td>
<td>fact</td>
<td>argument</td>
</tr>
<tr>
<td>10</td>
<td>right</td>
<td>benefit</td>
</tr>
<tr>
<td>11</td>
<td>issue</td>
<td>development</td>
</tr>
<tr>
<td>12</td>
<td>idea</td>
<td>essay</td>
</tr>
<tr>
<td>13</td>
<td>article</td>
<td>scheme</td>
</tr>
<tr>
<td>14</td>
<td>time</td>
<td>pro</td>
</tr>
<tr>
<td>15</td>
<td>question</td>
<td>right</td>
</tr>
<tr>
<td>16</td>
<td>side</td>
<td>point</td>
</tr>
<tr>
<td>17</td>
<td>effect</td>
<td>effect</td>
</tr>
<tr>
<td>18</td>
<td>evidence</td>
<td>area</td>
</tr>
<tr>
<td>19</td>
<td>reasoning</td>
<td>issue</td>
</tr>
<tr>
<td>20</td>
<td>result</td>
<td>time</td>
</tr>
</tbody>
</table>

In terms of the selection of the most frequent SN types, there is a fair degree of similarity in the two corpora. Both corpora share the following nine SNs in their top twenty: argument, problem, way, right, issue, result, example, time, effect. There are a number of words in the top-20 lists, however, that occur in one corpus but
not the other; the following eleven words only occur in the L1 English corpus: case, reason, claim, act, article, reasoning, side, fact, idea, question, evidence. On the other hand, the following 11 words only occur in the top-20 list of the learner corpus: advantage, disadvantage, method, ban, benefit, development, essay, scheme, pro, point, area. Particularly striking here is the very high usage of advantage and disadvantage and pro in the learner corpus, a point which will be returned to in the discussion.

Turning now to research question 4, concerning the differences in the range of different SN types, this feature is measured by comparing the frequencies of individual SN types in each of the corpora at different cut-off points. Table 4 shows this range across the two corpora. For the first two cut-off points — over 200 and over 150 occurrences — the Cantonese-speaking learners of English score higher than the native speakers (two to one and three to one respectively). After this, however, the L1 writers have a consistently wider range of SNs. As one goes down the list, in fact, a clear trend develops where the lower the cut-off point the greater are the superior numbers of SN types used by the L1 writers. This indicates a much wider overall range of different SN types used by the native speakers.

Table 4. Range of SN types across the two corpora

<table>
<thead>
<tr>
<th>Number of occurrences of items</th>
<th>Number of items occurring this many times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Native English</td>
</tr>
<tr>
<td>Over 200</td>
<td>1</td>
</tr>
<tr>
<td>Over 150</td>
<td>1</td>
</tr>
<tr>
<td>101–150</td>
<td>5</td>
</tr>
<tr>
<td>51–100</td>
<td>11</td>
</tr>
<tr>
<td>41–50</td>
<td>5</td>
</tr>
<tr>
<td>31–40</td>
<td>8</td>
</tr>
<tr>
<td>21–30</td>
<td>19</td>
</tr>
<tr>
<td>11–20</td>
<td>38</td>
</tr>
<tr>
<td>6–10</td>
<td>47</td>
</tr>
<tr>
<td>2–5</td>
<td>86</td>
</tr>
<tr>
<td>1</td>
<td>90</td>
</tr>
</tbody>
</table>

Finally, with regard to research question 5, concerning the relative frequencies of the different realisation patterns, we will consider those realisation patterns which learners are using less or more frequently than L1 English writers. Table 5 shows four patterns which the L1 writers are using much more frequently than the learners (numbers normalised per 100,000 words and raw numbers).
Table 5. Realisation patterns which the L1 writers are using much more frequently than the Cantonese-speaking learners

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Native English writers</th>
<th>Cantonese-speaking learners of English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per 100,000 words</td>
<td>Raw nos.</td>
</tr>
<tr>
<td>SN + of + np</td>
<td>232</td>
<td>256</td>
</tr>
<tr>
<td>SN + to + non-finite</td>
<td>195</td>
<td>215</td>
</tr>
<tr>
<td>SN + be + identifier</td>
<td>138</td>
<td>153</td>
</tr>
<tr>
<td>SN + that clause</td>
<td>119</td>
<td>131</td>
</tr>
</tbody>
</table>

As shown in Table 5, each of these patterns is being used over twice as many times by the L1 English writers (as compared to a 50% overall average — cf. Table 2). On the other hand, in spite of the overall higher frequency of use of the in-clause pattern by the L1 writers, there are some patterns which are being used more frequently by the learners. Table 6 shows the patterns favoured by learners.

Table 6. Realisation patterns which the learners are using much more frequently than the L1 writers

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Native English writers</th>
<th>Cantonese-speaking learners of English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per 100,000 words</td>
<td>Raw nos.</td>
</tr>
<tr>
<td>identifier + be + SN</td>
<td>128</td>
<td>142</td>
</tr>
<tr>
<td>SN + be + that + clause</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>SN + such as + np/V-ing</td>
<td>52</td>
<td>57</td>
</tr>
<tr>
<td>SN + of + V-ing</td>
<td>33</td>
<td>37</td>
</tr>
</tbody>
</table>

The significance of these differences in in-clause usage will be considered in the discussion section below.

5. Discussion

Given the hypothesised difficulties of the Cantonese-speaking learners of English with SNs, it is not surprising that the results reported above indicate that L1 writers make use of these nouns more frequently than do their L2 peers (research questions 1 and 2). It is important to stress the significance of considering the text-based frequencies in addition to the overall frequencies in this study, the former giving a more accurate — and less favourable — measure of learners’ ability to
use SNs. Indeed the different results of the aggregated and individual text-based frequencies may have methodological implications for corpus study in general, suggesting that aggregated statistics may mask internal variations in a corpus and even misrepresent the general tendency under the effect of outliers. Greater significance, it follows, should be attached to analysis on the individual text level within a corpus, especially when individual variations are assumed, as is often the case with learner corpora.

Returning to the study in hand, let us consider the results relating to research question 2, the relative frequency of the different functions. As demonstrated in the results, the frequencies of anaphoric across-clause SN and in-clause SN in the native-speakers’ corpus are over twice as high as those in the learner corpus. Given this finding, the comparatively much smaller difference in cataphoric across-clause SN between the two corpora — 45% more in the L1 writer corpus (although actually more frequent according to individual text-based frequencies) — is worth discussing.

A more qualitative look at the data relating to research question 2 may help to explain this discrepancy. One overall feature of the corpus is that the learner essays tend to be shorter (as noted, they were written under test conditions, unlike those of the native speakers). Given that cataphoric across-clause SNs often occur in introductory material, this may help in part to explain the learners’ preference for this pattern of SNs over other patterns. Additionally, there may be a teaching effect on the move structure of the essays and the lexico-grammatical realisations of such a structure, particularly in introductions, that encourage the use of certain patterns of SNs. Because these patterns and their associated nouns have been practised a lot, they are very familiar and learners may rely on them. Hasselgren (1994) refers to this phenomenon of learners relying on vocabulary patterns which are familiar as the ‘teddy bear’ effect. In an investigation into the ways Norwegian first year university students and upper sixth-formers cope with English vocabulary in their written language, Hasselgren found that these learners depend heavily on the familiar, either by “choosing words and phrases closely resembling their first language or those learnt early or widely used” (ibid.: abstract).

Take the following extracts (examples 16–20) from introductory paragraphs, where, in each case, SNs — reasons, pros and cons, arguments, functions, benefits, drawbacks and opinion — are used to introduce a series of reasons, pros and cons, arguments, etc. which will form the basis of the essay.

(16) Hong Kong’s education system always focuses on examination. Some bodies may say it is examination-driven. So, the marker must be fair to assess student’s standards. Because of lots of reasons, I do not agree peer assessment.
(17) Here, let us try to find out the *pros and cons* of developing Hong Kong’s country park.

(18) In this essay, I examine the *arguments* which support importing professionals from Mainland China into Hong Kong and those which do not.

(19) The tourism sites in Hong Kong can be classified into four main categories — cultural, natural, man-made and historical. Each of them has different *functions* for tourists.

(20) In this article, I will discuss both *benefits* such as reducing cost, preventing pollution problem and *drawbacks* like huge capital input is need and the market demand for recycled products. Also, I will provide my personal *opinion* on recycling in Hong Kong.

Greater use of the cataphoric pattern than other patterns by the learners may also be the result of the text type of many essays, which follows a “compare and contrast”, “advantages and disadvantages” pattern. Indeed, as the following extracts (examples 21–23) from opening paragraphs indicate, a favourite introductory pattern is to point forward to *advantages* and *disadvantages*.

(21) In this essay, I discuss the *advantages* and *disadvantages* of students using credit cards.

(22) In this essay I make an effort to discuss the *advantages* and *disadvantages* of banning smoking in restaurants.

(23) In this essay, I will elaborate on both *advantages* and *disadvantages* and then I will point out whether it is beneficial to use recycling.

The following extracts (24–28) from a D+ essay show that, while overall, as demonstrated by Flowerdew (2006), better writers use more SNs than do less competent ones, even at a low level, learners may make use of a favourite SN or nouns (*advantage* and *disadvantage* again), probably due to the teaching and/or text type effect. Such formulaic language was not found in the native-speaker corpus:

(24) … This essay is going discuss the *advantage* and *disadvantage* of importing professional i.e. qualified professionals, technical staff, administrators and managerial personnel from Mainland China into Hong Kong.

(25) The first *advantage* of importing professionl from Mainland China is that …

(26) The second *advantage* is that …

(27) However, there are serval *disadvantage* of the imported professionals.
Evaluating the evidence of this essay, there are several advantage and disadvantage of imported professionals from Mainland China ...

More globally, if we take the pair of SNs advantage(s) and disadvantage(s), it is found that this pair is used disproportionately in the learner corpus. Indeed, as shown in Table 3, advantage and disadvantage are the second and third most frequent SNs respectively in the entire learner corpus, while neither is in the top 20 for the L1 corpus. Also in the top 20 for the learner corpus are the near synonyms of advantage, pro, as in pros and cons, and benefit. If this over-use of advantage, disadvantage, pro and benefit is taken into account with regard to research question 1 concerning the overall frequency of SNs in the two corpora, with the 17% greater use by the L1 writers, then, by excluding the numbers for these four items, the 17% greater use is increased to 32% (41 per thousand vs. 31 per thousand).

Overall, it can be said that, given the shorter nature of the learner essays, the possibility of a teaching effect, and a particular text type favoured in some of the essays in the corpus, the data reported here suggest that the students may have been encouraged to use more SNs than they might have done under different conditions. The frequency data reported in this study, therefore, may give an exaggerated impression of the learners’ ability to use across-clause SNs.

Turning now to the in-clause usage, the greater usage by the L1 English writers (17 vs. 12 per thousand, see Table 6) is more predictable. Here, there is no obvious teaching effect and the syntactic complexity of the patterns associated with this type would seem to be a likely possible explanation of its greater use by the L1 writers, learners likely preferring more congruent syntactic patterns and ones which correspond more closely to Chinese information structure. Thus, for example, rather than the complex nominalisation found in example (8) in the list of patterns presented under the heading of research question 3 above — a consideration of the sanctity and value of human life — a learner might well prefer something like *if you consider how holy and valuable human life is. This might be because in Chinese the complex post-modifier in English would have to be transposed as a pre-modifier and would sound extremely clumsy, as in example (29):

(29) 对人生的神圣与价值的思考
Dui Ren-sheng De Shen-sheng Yu Jia-zhi De Si-kao
“of the sanctity and value of human life consideration”

A similar explanation might be given for the pattern SN + to + non-finite. For instance, in example 10 above, a firm decision to heed or to disregard the pleas of opponents to corporal punishment, the SN pattern decision to might be written more congruently as when somebody decides to.... Similar arguments of congruence might be put forward for patterns 3 (SN + be + identifier) and 4 (SN + that clause) in the list of patterns presented in Table 5.
As regards research question 3, concerning the differences in the selection of specific nouns, the results have shown a fair degree of conformity, with nine of the top 20 SNs in both corpora being shared. Of the nouns which are not shared, the most obvious possible reasons for this are the text types and topics of the essays in the two corpora. The high frequencies of *advantage*, *disadvantage*, *pro* and *benefit* have already been posited as a possible result of one of the text types preferred in the learner corpus. Another high-frequency item in the learner corpus, *ban*, is also likely an artefact of the data, one of the essay topics in the learner corpora concerning the issue of the banning of smoking. The items in the top-20 list for the L1 English writer corpus, on the other hand, do not seem to be related to any specific text types or topics. Broadly speaking, however, if the possible text type- and topic-related nouns from the learner corpus are set aside, there is a good deal of overlap in the overall choice of SNs from among those most frequently used, suggesting that the learners have a basic competence in the most important ones.

Turning now to research question 4, regarding the differences in the range of different SN types, it turned out that the range of SNs is clearly much greater for the L1 writers, as would be expected, with the learners relying on a smaller set which they use more frequently, as suggested by the finding for research question 3, where learners used a similar range of more frequent SNs to those used by the L1 writers. This might partially be explained by Hasselgren’s (1994) ‘teddy bear’ effect, as referred to earlier, with learners depending on those nouns and patterns they are most familiar with.

Regarding the final research question, question 5, concerning the relative frequencies of the different realisation patterns of the in-clause function, the tendency for the L1 writers to use certain patterns more than the learners (cf. Table 5) might be explained, in part, of course, as a function of the overall greater usage of in-clause patterns on the part of these L1 writers (cf. research question 2). But how is one to explain the fact that, in spite of an overall inferior usage of the in-clause patterns, certain of these are nevertheless made more use of by the learners than by the L1 writers (Table 6)?

First of all, less usage by the L1 writers in all cases might be affected by their competence with a wider range of patterns overall. In other words, use of a given pattern might be less frequent because of an overall wider range of patterns being used. This is likely given the wider range of SNs used by the L1 writers (cf. Table 4). Secondly, the frequency of a given pattern may be affected by the particular SNs which take that pattern. For example, the most frequent SN in the L1 English writer corpus is *argument*. This noun follows the patterns SN + that clause (e.g. *the opponents’ arguments that nuclear power requires a huge capital investment*) and SN + be + that (e.g. *Their argument is that there is just so little of it out there*). *Case*, on the other hand, the third most frequent SN used by the L1 English
writers, prefers the pattern SN + of + np (e.g. *the case of white males experiencing reverse discrimination*). The fact that these high frequency nouns take particular patterns may affect the overall numbers. It is worth noting that, although there were sometimes colligational errors in the use of nouns by the learners, as reported by Flowerdew (2006), this did not extend to a systematic use of a wrong pattern with a given SN.

Additionally, possible explanations might be sought individually for each of the patterns concerned. In the case of identifier + be + SN, as in *Students hold credit card as a mean of payments is a trend in universities*, this usage is possibly related to transfer from Cantonese, which is a topic prominent language, i.e. the topic is typically mentioned first in the clause and is followed by the comment (Li & Thompson 1976). This topic prominent structure might be selected where English might prefer something like *A trend in universities is for students to hold credit cards as a means of payment*. Evidence to support this interpretation is provided by the fact that these patterns, which sound inappropriate in English, translate quite naturally into Chinese.

Regarding SN + be + that clause, this would seem to be quite a straightforward pattern that translates naturally from Chinese, so it might be favoured because of its relative simplicity for a Chinese learner. Also, a possible teaching effect cannot be ruled out. A teacher, in focussing on SNs and how to use them, might well favour this as a stereotypical pattern on the basis of its common appearance in grammar and course books. Similarly SN + such as + np/v-ing translates naturally from Chinese, e.g. example (30). SN + of + V-ing, on the other hand, would have the post-modifier in English as a pre-modifier, so there is no obvious reason for this preference, e.g. example (31):

(30) 好处, 例如降低成本、防止污染问题
Hao-chu, Li-ru Jiang-di Cheng-ben, Fang-zhi Wu-ran Wen-ti
“benefits such as reducing cost, preventing pollution problem”

(31) 禁烟的好处
Jin Yan De Hao-chu
“of banning smoking benefits”

6. Conclusion

Flowerdew (2006) demonstrated that frequency (and accuracy) of use of SNs is a developmental phenomenon which increases with overall writing ability. As would be predicted, therefore, this paper demonstrated that Cantonese-speaking learners of English use less SNs in terms of overall frequency and in terms of the range of
use than do their L1 peers. In addition, more detailed findings have suggested that the stereotypical use on the part of the learners of certain cataphoric across-clause SNs might be due to a training or text type effect. This was an additional finding, not anticipated by the research questions.

What this implies for pedagogy would seem to be rather straightforward: that learners, on the one hand, need to increase their frequency and range of use of SNs, while teachers, on the other hand, need to guard against possible teaching and text type effects, by introducing a wider range of nouns and working with a wider set of text types. On the other hand, it is also worth pointing out that, for lower level learners, a small range of SNs can serve as a useful, if predictable and repetitive, cohesive devise, allowing such learners to scaffold connected text in the earlier stages of their writing development.

Teachers therefore need to balance the value of a small range of devices which allows students to write coherent, if non-L1–like texts, and the need to develop a wider range of SNs, if learners are to approximate more closely to L1 usage. In more general terms, in showing the much greater frequency and range of use of signalling nouns on the part of L1 writers as compared to the learners, this study confirms the importance of signalling nouns as an essential dimension of (lexical) cohesion, contributing to the overall coherence of a text.

Notes

* I would like to acknowledge the help of Delian Gaskell and Sylvia Xiao Chen in the preparation of this paper. I would also like to acknowledge the helpful comments of two anonymous reviewers and Michaela Mahlberg in revising an earlier version of the paper. This paper was funded as part of a Hong Kong University Grants Committee Competitive Earmarked Research Grant #9040857.

1. Examples (1)–(5) are taken from the corpus used in Flowerdew (in preparation).

2. Abstract nouns may also refer exophorically, i.e. their meaning is to be found outside the text. In such cases they are not counted as signalling nouns.

3. The essays were graded on a scale from A to F, where A is considered to be excellent, F is a fail, D is considered less than satisfactory, and D+ is a score mid-way between D and C.

References

Flowerdew, J. In preparation. “SNs in academic genres and registers”.

All rights reserved


Winter, E. O. 1992. "The notion of unspecific versus specific as one way of analysing the information of a fund-raising letter". In W. Mann & S. A. Thompson (Eds.), Discourse Description:

Author's address

John Flowerdew
School of Education
University of Leeds
Hillary Place
Leeds LS2 9JT
UK
J.L.Flowerdew@education.leeds.ac.uk