Combining intuition with corpus linguistic analysis: a study of marked lexical chunks in four Chinese students’ undergraduate assignments

How to cite:

For guidance on citations see FAQs

© 2012 The Author
Version: Accepted Manuscript
Link(s) to article on publisher’s website:
http://ojs.ub.gu.se/ojs/index.php/njes/article/view/1609/1411

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
Combining intuition with corpus linguistic analysis:
A study of marked lexical chunks in four Chinese
students’ undergraduate assignments

Maria Leedham, The Open University, UK

Abstract
In the literature on lexical chunks, a dichotomy is frequently implied
between intuition-based methods of finding language ‘formulaic’ and
frequency-based means of extracting ‘n-grams’. In this paper, a case
study of four Chinese students’ undergraduate assignments is described
in terms of marked or atypical lexical chunks revealed through close
reading and those found through keyword analysis, when compared with
a reference corpus of similar writing by British undergraduates. The
paper discusses the benefits of combining the two approaches, arguing
that this gives clearer insights into the personal phraseological profiles of
the students’ writing than either can offer alone.

1. Introduction
More and more Chinese people are choosing to study abroad, with
284,700 doing so in 2010 (British Council, 2012); this study is
increasingly taking place at degree level in English-speaking countries.
Despite this growth, comparatively little research has been carried out on
Chinese students’ assessed undergraduate writing, with most studies
exploring either short texts or longer, Master’s level theses (e.g. Chuang
and Nesi, 2006; Hyland, 2008). This study takes a case study approach in
focusing on the writing of four Chinese students in UK Higher
Education; their assignments are compared with texts in the same
disciplines, and also with larger corpora of L1 (first language) Chinese
and L1 English student texts¹ to uncover features of the language which
are particular to the individual, the discipline, and the L1. It should be
noted that the L1 English writing is not intended to be normative. Both
the L1 Chinese and L1 English texts used in the study are successful

¹ Note that ‘assignment’ and ‘text’ are used interchangeably in this paper.
assignments and were awarded a IIi or I in the UK system (equivalent to ‘merit’ or ‘distinction’). Moreover, it is recognized that L1 English undergraduate students are also novices in learning the conventions of academic writing within their discipline and as such are not necessarily ‘better’ academic writers.

The comparisons are carried out in terms of the nature of the ‘lexical chunks’ or ‘chunks’ used in the writing; chunks are used here as an umbrella term to cover frequently-occurring sequences of words and collocations or words which ‘predict one another, in the sense that where we find one, we can expect to find the other’ (Durrant, 2008: 5). Research into the contribution made by lexical chunks to academic writing has proliferated in recent years as these are widely regarded as indicators of competent language use (e.g. Ådel and Erman, 2012; Biber and Barbieri, 2007; Cortes, 2004; Hyland, 2008). Using preferred, conventionalized ways of expressing meaning is easier for the writer since ideas can be expressed using prefabricated units rather than being constructed anew. It is also easier for the reader since existing phrases are more easily recognized than novel ones (cf. Wray and Perkins, 2000). Learning to write in academia can thus be viewed as using chunks which the reader recognizes as particular to the discipline and which therefore help to establish the writer’s membership within the disciplinary community (e.g. Li and Schmitt, 2009).

This study examines those chunks which are marked or atypical in four Chinese students’ writing when compared with a larger corpus of writing in the same discipline or with a corpus of L1 English student writing. The term ‘marked’ is employed here in the sense that the chunks appear unusual in the context of academic writing, perhaps due to their informality or to their idiosyncratic nature. The study is thus different to the majority of corpus studies which concentrate on high frequency items meeting a minimum dispersion level across individuals and texts and which remove any idiosyncratic chunks (e.g. as in Chen and Baker’s, 2010, study of four-word lexical chunks in Chinese students’ writing). In this study, on the other hand, rare chunks are of interest since these can reveal unusual and hence noticeable aspects of individual student writing. In this, the paper draws on corpus stylistics work on exploring the work of individual writers in order to raise awareness of distinctive features of the writing (e.g. Coniam, 2004; Lee and Swales, 2006).

This paper reports on findings from the study’s two objectives: the
first of these is to describe features of Chinese students’ written English assignments; the second aim is to contrast two approaches to identifying lexical chunks and compare what is revealed through each method. In the first method, each student’s assignments are read by the author in order to identify salient lexical chunks, that is, those which appear to be marked or atypical in some way and which may be idiosyncratic to the individual or L1 group. Using WordSmith Tools (v. 5; Scott, 2011), the number of occurrences of each identified chunk is then found within all texts by the same student, and is compared with the number found in reference corpora of L1 English assignments from the same discipline and also from a larger corpus of L1 Chinese undergraduate assignments. The second method begins from corpora, using WordSmith Tools to identify keywords in each student’s writing using the same reference corpora as the first method. The co-text of the chunks uncovered through each method is then explored and the chunks are grouped into categories. Discussion in the paper centres on the benefits of using reader intuition and corpus tools as the means of initially identifying lexical chunks which are marked in an individual’s writing, or salient in a discipline or L1 grouping.

Section 2 describes the two methods more fully. This is followed by a description of the data (section 3), findings and discussion from each method (section 4) and conclusions.

2. Two methods of identifying and extracting lexical chunks

Wray (2008: 93) discusses an inherent circularity in identifying lexical chunks, since ‘you cannot reliably identify something unless you can define it’, yet in order to define it, you must have some examples to study. A theorist’s underlying view of chunks is therefore bound up with the choice of identification method; for example defining chunks by how many times they occur leads to a computational method of identification, excepting very small samples where counts can be manual (see Wray, 2002, for discussion of different methods of identification and extraction). In this paper I suggest that a major division between types of lexical chunk hinges on semantic unity, as this points to the divide between chunks as intuitively-determined, psychologically ‘complete’ linguistic items, and chunks as frequently-occurring, well-dispersed phenomena. For example, a lexical chunk occurring just once in a corpus
(a hapax legomenon) may be semantically ‘whole’ but would not be captured through a frequency-based search. Conversely, a chunk can occur frequently but not feel semantically ‘complete’ (e.g. *that there is a*). The criterion of frequency is the primary defining feature of chunks known variously as ‘clusters’ (e.g. Scott, 2011), ‘n-grams’ (e.g. Milton, 1999), and ‘lexical bundles’ (e.g. Biber et al., 1999); these require parameters to be set for the length of the chunk, threshold for minimum frequency, and the minimum number of texts for dispersion in order to avoid idiosyncrasies and also repetitions due to localized topics. For example, for Biber et al. (1999) four-word lexical bundles must occur ten or more times in a corpus and across a minimum of five texts per register to qualify as bundles. A way of verifying the holistic validity of chunks retrieved through frequency is to apply a statistical measure of collocation such as the Mutual Information (MI) test. This test measures the extent to which the observed frequency of co-occurrence differs from what might be (statistically) expected, that is, the strength of association between words. MI works less well with very low frequencies, however, and in these cases the t-score is a more reliable measure since this takes raw frequencies of occurrence into account.

Within the umbrella concept of a ‘lexical chunk’, I adopt two commonly-used terms. ‘Formulaic sequence’ is now widely-used to refer to the intuitively identified chunk, defined by Wray (2002: 9) as ‘a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated’. The ‘n-gram’ (and thus ‘3-gram’, ‘4-gram’) is a chunk defined by frequency of occurrence and which therefore may or may not be semantically whole.

Figure 1 illustrates how these labels fit within other commonly-used terms in the literature. The left-hand circle represents formulaic sequences and the right-hand one shows n-grams. Within the overlap of the two circles are examples of chunks which are both frequently-occurring and semantically-whole, such as frequent connectors (e.g. *on the other hand*). In the left-hand circle but overlapping slightly with the right-hand one are Moon’s (1998) Fixed Expressions and Idioms (FEIs) (e.g. *kith and kin*); these can be frequent or infrequent, but are all contained within the circle of semantically-unified formulaic sequences.

---

2 See discussion of MI and t-score tests on the Collins Wordbank site here: [http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html](http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html).
Figure 1. Lexical chunks
Also within the left-hand circle and overlapping with n-grams are semantic sequences (Hunston, 2008), shown here within a dotted circle to indicate the abstracted and thereby permeable nature of these chunks. Semantic sequences are incomplete structures, requiring lexis to instantiate each example and subsume the category of collocational frameworks (Renouf and Sinclair, 1991) for example ‘a + noun-classifier + of + noun-category’ instantiated as a kind of experiment. The subsumed collocational framework in this case is a * of, giving rise to a kind of, a form of.

In the right-hand circle of Figure 1 but overlapping with formulaic language are categories of frequently-found n-grams as these may or may not be semantically whole units; here, 2-grams are shown as contained within 3-grams, and so on (e.g. on the within on the other which is in turn within on the other hand). Solely in the n-gram circle are those chunks which are frequently occurring but which are not semantically whole units (e.g. the other hand the).

The next two subsections describe the methods used in the study to find lexical chunks.

2.1 Finding formulaic sequences through intuitive reading
The use of intuition to manually extract formulaic sequences from the writing of others entails consideration of issues such as inter-rater reliability, within-rater consistency, and decisions as to where to place sequence boundaries. Moreover where the rater has a different L1, they may be unable to determine chunks which are valid for the writer/speaker (Foster, 2001). Thus, the formulaic sequences identified may vary significantly in quality and quantity if raters are linguistically-aware discipline specialists possessing familiarity with the writer’s L1, compared to raters without this knowledge. However, providing specific guidelines as to the boundaries of chunks would reduce the freedom of an individual’s intuition and impose the researcher’s views. Despite the inherent difficulties in the intuitive identification of chunks, many studies rely on intuition at some level, whether for the initial extraction of chunks or to refine a computationally-produced list of chunks (e.g. Baigent, 2005; Leedham, 2006; Li and Schmitt, 2009; Nesselhauf, 2005; Schmitt et al., 2004; Wray and Namba, 2003).
Some of the issues discussed above are avoided if, rather than multiple raters, a single rater is used to identify chunks (cf. studies using single raters carried out by Baigent, 2005; Nesselhauf, 2005). Moreover, the rater-analyst is likely to spend far longer on the laborious task of reading and rereading texts in order to identify sequences. For this study, the overall size of contributions from the case study students meant that it was not viable to ask other people to identify sequences (the texts total over 48,000 words from L1 Chinese students alone). Instead I employed my intuition as an applied linguistics researcher with 20 years’ experience of teaching English for Academic Purposes and particular familiarity with Chinese students’ writing styles. This experience gives me some insight into common features of the writing of this group of students, though may also mean I fail to observe language which may be salient to other readers. Checks were made to ensure the identified chunks were in fact marked by asking two similarly-experienced English language tutors to confirm the sequences as unusual in academic writing.

I first carefully read all assignments by each case study student in conjunction with assignments from L1 English students in the same discipline. Formulaic sequences were identified which were salient because of their apparent atypicality within academic writing, or because they appeared to be favoured by the particular student (cf. Wray and Namba’s, 2003, list of possible criteria for pinning down intuitive judgements). Following this, I used WordSmith Tools to determine the frequency of each identified sequence within all assignments from the same student, and also searched reference corpora of texts in the same discipline and from each L1 group. Log likelihood tests were carried out where there were sufficient raw examples. These searches enabled me to establish in each case whether, based on the (albeit limited) data, the chunk appears to be idiosyncratic within the writing of a single student, or is frequent within the particular discipline or L1 grouping. I achieved a measure of reliability through carrying out the process twice, with an interval of six months in between. The second close reading of the assignments revealed additional idiosyncratic sequences, suggesting that the more time spent on this task the greater the number of sequences found (cf. Leedham, 2006).
2.2 Finding keywords though WordSmith Tools
Unlike intuitive reading, n-grams searches do not rely on knowledge of the discipline content of texts or familiarity with the writing of the student group. However, the use of corpus linguistic tools still involves human decisions as to the search parameters used (the length of the chunk, the minimum frequency of occurrence, and the dispersion of texts it must be found in). These essentially arbitrary judgments are often carried out according to the pragmatic measure of how many chunks are generated under a particular group of settings. Too few chunks would result in insufficient data to analyze, too many may overwhelm the researcher and make it hard to assess the results (Schmitt et al., 2004).

In this study, each student’s texts comprise a small corpus while the L1 English texts from the same discipline area form a corresponding reference corpus. N-grams were extracted based on keyness (using the log likelihood test) in line with many previous studies of lexical chunks in written language (e.g. Biber et al., 1999; Hyland, 2008; Schmitt et al., 2004). ‘Key’ items are those which occur statistically more often in a small corpus than a larger reference corpus, relative to the total number of words in each corpus, meaning that keyness is thus a ‘matter of being statistically unusual relative to some norm’ (Culpeper, 2009: 34). Using WordSmith Tools (with the setting $p=0.00001$), I searched for all keywords of two words or longer. The log likelihood test was selected to determine keyness, following Dunning’s (1993) argument that chi square and mutual information tests are less valid than the log likelihood ($G^2$) test where counts are low. Any keywords subsumed within longer ones were removed.

2.3 Comparison of methods
Table 1 summarizes the pros and cons of human intuition versus corpus tools as methods for finding lexical chunks.

3. The data
This section contains an overview of the four students and their individual contributions to the corpus, then gives details of the reference corpora.
<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Using human intuition to find formulaic sequences</th>
<th>Using corpus tools to extract n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Sequences do not cross clausal boundaries.</td>
<td>• Ngrams frequently occur across clausal boundaries.</td>
</tr>
<tr>
<td></td>
<td>• Sequences are psychologically real and stored as wholes in the mental lexicon (Schmitt et al., 2004).</td>
<td>• There is evidence to suggest that not all bundles are stored as wholes in the mental lexicon (Scott, 2011).</td>
</tr>
<tr>
<td>Pros</td>
<td>• Chunks found will feel ‘whole’.</td>
<td>• Large quantities of data can be analyzed quickly and accurately (as far as tagging and software allow).</td>
</tr>
<tr>
<td></td>
<td>• They are thus ‘teachable’.</td>
<td>• Findings are easily replicable.</td>
</tr>
<tr>
<td></td>
<td>• Single instances of a chunk can be identified.</td>
<td>• Patterns that are not salient to the human reader are revealed.</td>
</tr>
<tr>
<td>Cons</td>
<td>• Only relatively small quantities of data can be analyzed.</td>
<td>• Representativeness is only as good as the corpus compilation</td>
</tr>
<tr>
<td></td>
<td>• Very timeconsuming.</td>
<td>• Ngrams cross clausal boundaries and may feel unnatural.</td>
</tr>
<tr>
<td></td>
<td>• Inconsistent results – the longer you look, the more chunks you find (Leedham, 2006).</td>
<td>• Many ngrams may not be readily usable within teaching materials.</td>
</tr>
<tr>
<td></td>
<td>• Tendency to find what you expect to occur in the data.</td>
<td>• Chunks occurring once only in the corpus are missed.</td>
</tr>
<tr>
<td></td>
<td>• Different people have different intuitions, depending on their linguistic exposure (Hoey, 2005). E.g. a NS may not notice L2 English students’ chunks in English.</td>
<td>• Corpus tools cannot distinguish between language used in a formulaic way and the same language which is built up e.g. keep your hair on can be metaphorical or literal (do not remove your wig) (Wray, 2002: 31).</td>
</tr>
<tr>
<td></td>
<td>• Discrepancies within one individual’s categorizations. (Foster, 2001).</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Hard to replicate findings.</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparing the use of human intuition and corpus tools to find chunks
3.1 The students

The data in this study was taken from the British Academic Written English (BAWE) corpus; this reflects the situation within the UK as a whole in that Chinese students are the largest L2 English student group (British Council, 2012) (see Nesi and Gardner, 2012, for details of BAWE corpus compilation). Four student contributors fulfilled the criteria set for this case study; these were having Chinese (Mandarin or Cantonese) as an L1, undertaking all secondary education in their home country, and submitting assignments to the corpus from years 1/2 and year 3 of undergraduate study. All four students, two males and two females, were in their early 20s during their (full-time) degree courses. Pseudonyms are used throughout. In total, there are 29 assignments comprising 48,367 words from the four students in this study (Table 2).

Table 2: Wordcounts and number of texts per student

<table>
<thead>
<tr>
<th>Student (gender) (BAWE ID)</th>
<th>Degree discipline</th>
<th>No. words in year 1¹</th>
<th>No. words in year 2</th>
<th>No. words in year 3</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wei (m) (0254)</td>
<td>Engineering</td>
<td>3,084 (3)</td>
<td>6,347 (4)</td>
<td>3,348 (3)</td>
<td>12,779 (10)</td>
</tr>
<tr>
<td>Feng (f) (6008)</td>
<td>Food Science</td>
<td>(none)</td>
<td>4,513 (5)</td>
<td>9,170 (5)</td>
<td>13,683 (10)</td>
</tr>
<tr>
<td>Mei-Xie (f) (3018)</td>
<td>HLTM*</td>
<td>4,462 (2)</td>
<td>5,047 (2)</td>
<td>3,859 (1)</td>
<td>13,368 (5)</td>
</tr>
<tr>
<td>Hong (m) (3085)</td>
<td>HLTM</td>
<td>3,143 (1)</td>
<td>2,581 (1)</td>
<td>2,813 (2)</td>
<td>8,537 (4)</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>10,689 (6)</td>
<td>18,488 (12)</td>
<td>19,190 (11)</td>
<td>48,367 (29)</td>
</tr>
</tbody>
</table>

*HLTM = Hospitality, Leisure and Tourism Management

Further texts from L1 Chinese students and from L1 English Engineering; Hospitality, Leisure and Tourism Management (HLTM) and Food Science and from a similar range of genres (such as essays, laboratory reports and case studies) are used as reference corpora in this study. These total 279,695 for the Chinese reference corpus and

¹ Information within parentheses refers to number of texts.
1,335,676 words for the English one (Table 3). The discipline subcorpora (e.g. English-Engineering) are a subset of the texts within the L1 English reference corpus (Eng123).

Table 3: Wordcounts and number of texts for reference corpora

<table>
<thead>
<tr>
<th>Corpus name (L1 + discipline)</th>
<th>No. of Texts</th>
<th>Word counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Engineering</td>
<td>97</td>
<td>203,379</td>
</tr>
<tr>
<td>English-Food</td>
<td>28</td>
<td>73,402</td>
</tr>
<tr>
<td>English-HLTM</td>
<td>55</td>
<td>64,563</td>
</tr>
<tr>
<td>Chi123</td>
<td>146</td>
<td>279,695</td>
</tr>
<tr>
<td>Eng123</td>
<td>611</td>
<td>1,335,676</td>
</tr>
</tbody>
</table>

4. Findings and discussion
This section discusses the findings from each of the two methods of extracting lexical chunks.

4.1 Findings from intuitive reading plus corpus searches
In this section normalized figures per one million words (pmw) are given to facilitate comparison between differently-sized corpora. Findings are discussed under thematic headings.

Idiosyncratic sequences
Sequences in this group are those which were marked on reading through an individual’s assignments, yet were found through concordance searches to occur infrequently in the larger corpora of the same discipline or L1 groupings, that is, they are idiosyncratic to the individual concerned. It should also be noted here that this investigation begins
from the writing of four individual L1 Chinese students; if four L1 English students were taken as case studies, then equally idiosyncratic chunks particular to these individuals might be found.

The chunk *in light of this* appeared marked on reading Mei-Xie’s texts, and a corpus search showed this linking chunk occurs just 3 times in a single assignment from Mei-Xie and only once more in Chi123, for example:

(1) …the stock market is at or near a temporary peak. *In light of this*, it can be suggested that…

(2) …is room for market capitalisation growth of IHG. *In light of this*, it is recommended that buying IHG…

(Mei-Xie)

There were only 5 occurrences of this sequence in Eng123 (1.3 million words), all in clause-initial position and demarcated by a comma, though a similar chunk, *in the light of* (followed by a noun phrase), was more prevalent in this L1 English corpus with 11 occurrences (2 in Chi123).

Similarly, the sequence *in one word* is noticeable in assignments written by Wei, an Engineering student. This chunk is used twice, in both cases to summarize a previous section:

(3) …one again originally. *In one word* computer based tools contribute…

(4) …placement sensors. *In one word* the overall system can be described…

(Wei)

A search in Chi123 reveals just three additional instances of this sequence; there are no occurrences in Eng123.

A further sentence-initial connecting chunk is used by two of the case study students yet is still infrequent in the reference corpora. Feng and Mei-Xie use the sequence *that is why* to signal an explanation of a phenomenon. Two other L1 Chinese students together account for three uses of this chunk, making a total of seven occurrences in Chi123 (Figure 2, lines 1-7) and just five in Eng123 (Figure 2, lines 8-12) giving
a significance figure of \( p = .01^4 \).

\[ p < 0.05; \text{critical value} = 3.84; \]  
\[ ** p < 0.01; \text{critical value} = 6.63 \]  
\[ *** p < 0.001; \text{critical value} = 10.83; \]  
\[ **** p < 0.0001; \text{critical value} = 15.13 \]

\[ N \text{Concordance} \]

1 price compared with a perfectly competitive industry. That is why monopoly is less efficient. Monopoly is a
2 has a noticeable effect on the viscosity of the liquid. That is why cream (38% fat) is thicker than milk
3 are neglected which leads to poor service quality. That is why Visser (1991) suggests formality is a
4 real way, and the authenticities are very harmonious. That is why Emrol Morris works are almost received
5 immigrants is the best way of solving the problems. That is why I think the racism will be disappeared in
6 3 & 5 didn't take effects of pre-tilt into account. That is why the relationship of Equation 5 should be
7 issue for deciding which food products to purchase. That is why sensory analysis is vital to evaluate and
8 and put on the shelf it can be less than a week. That is why people are starting to prefer the
9 is ever changing and no two jobs are ever the same. That is why it is of high importance that I review my
10 and the other who could have committed the crime. That is why in many situations the statements of
11 to admit, reacting to basic needs and stimuli. Maybe that is why it was conceived as a science and the
12 is the fact that no cost information is displayed. That is why it is important to calculated measures

Figure 2: \textit{that is why} in Chi123 and Eng123

Many of the idiosyncratic sequences identified seem a little incongruous with the generally formal style of the assignments. For example, Hong and Mei-Xie’s writing includes the only three nominalized instances of \textit{must} in Chi123; there are just two occurrences in Eng123:

(5) … but simply writing a responsible tourism policy is no longer enough. It is \textit{a must} to show practical action, so that the tourism destinations can… (Hong)

(6) Besides enjoying the benefits the designation offer, it is \textit{a must} for Marriott Liverpool City Centre Hotel to bear the responsibility… (Hong)

(7) On the contrary, prior similar industry experience is not \textit{a must} since training will be provided. (Mei-Xie)

This chunk has perhaps been acquired through these Hospitality students

\footnote{4 Using Rayson’s log likelihood calculator \( \text{http://ucrel.lancs.ac.uk/llwizard.html} \)}
reading tourism brochures or job adverts and then appropriating the item within their academic writing. The similarly informal chunk *get rid of* is salient in Hong’s writing, yet occurs just twice in Chi123 overall:

(8) …a winning city, the authorities of Liverpool have to rebuild its image to *get rid of* the negative picture. (Hong)

(9) To have more accurate results, methods to *get rid of* RNase should be included. (Biology, Chinese student)

The final sequence discussed in this section is not salient due to any mismatch of formality, but is simply an unusual adaptation. It occurs just once in the corpora in Hong’s HLTM writing in the context of a report on how the Scottish tourist board can improve their tourism figures:

(10) …and boost its marketing campaigns in order to *catch the world’s eyes* on Scotland. (Hong)

This creative adaptation of the idiom *to catch someone’s eye* can be viewed as taking ownership of the language, rather than merely using whole idioms in their original form. Creativity in language, argues Hoey (2005: 53), comes from ‘the way we select from a lexical item’s primings and from our ability to ignore some (though rarely all) of these primings’. L2 English writers may have what Hoey terms ‘incomplete primings’ in comparison with L1 English writers since they lack the colligational and collocational knowledge which comes from sufficient quantity of input. However this should not exclude the majority of the world’s English speakers from creatively manipulating language (cf. Prodromou’s, 2007, argument for wider acceptance of L2 English writers’ and speakers’ innovations or *creative idiomaticities*).

The fact that the examples in this section are salient to this reader, yet infrequently used, illustrates the usefulness of corpus searches as a checking mechanism. A writing tutor or other reader may notice unusual uses of language and form the impression that particular chunks are widespread in the writing of an individual or an L1 group. Sequences in the following sections, in contrast, were found to occur more widely than in the four case study students’ writing; thus the case study examples provide a way in to wider analysis.
Vague and informal sequences

While a degree of vagueness can be appropriate as it avoids the stiltedness of over-specification (Channell, 1994), the expressions considered in this section seem to be employed out of context as they are more commonly associated with speech. 'Informal' is used here to refer to chunks which appear less appropriate in the context of academic writing. All chunks were checked in Biber et al. (1999) and also with the two additional raters to confirm that they were more informal than might be expected in academic writing.

The first sequence to be considered is *more or less*, found initially in Hong's writing:

(11) In catering services, restaurants in Oxford and Bath are *more or less* the same. (Hong)

On checking the corpora, I found nine instances of this chunk in Chi123 (Figure 3, lines 1-9) and six instances in Eng123 (lines 10-15), a significant difference at \(p=.001\).

While one sense of *more or less* in Figure 3 can be unpacked to mean *more X or less X* (e.g. line 11 allowing more air or less air through), most lines use *more or less* as a whole chunk meaning ‘approximately’ and appear incongruent with the otherwise formal text. The use of *than* following *more or less* is hard to process (more or less than what?), even
viewed with greater context.

The vague sequence *a little bit* was observed in three of the case study students’ writing, for example:

(12) At that time, I found that this hotel is *a little bit* out of my expectation. (Hong)

Lines 1-8 in Figure 4 show all occurrences of *a little bit* in Chi123, and lines 9-10 the only 2 occurrences in Eng123 (significant to $p=.0001$).

A search for *bit* in both Chi123 and Eng123 (with the removal of references to a computer *bit*) produced 21 and 23 instances respectively from a wide range of disciplines and genre families (significantly more frequent in Chi123, $p=.0001$). A collocate search suggests that the most common chunk for both student groups is *a bit* followed by an adjective e.g. *a bit extreme/high/more difficult/technical/wetter*. The L1 English students also use the pattern *a bit of a + N*, e.g. *a bit of a victim, a bit of an issue, a bit of a dog’s breakfast* (though the intriguing final example is a newspaper quotation, cited in a Law essay). This pattern occurs mainly in reflective sections of assignments, where informal language seems more acceptable. For example:

(13) The conclusion was also *a bit of a victim* in my editings, bringing it down to one small sentence for each of the areas of discussion. (L1 English, Cybernetics)

Thus, the L1 Chinese students make greater use of *bit* and use this across
more formally-written texts. The conversational nature of *bit* is confirmed by Simpson-Vlach and Ellis’ (2010) extraction of ‘academic formulas’ in which *little bit about* and *talk a little bit feature* in the list of *spoken* academic formulas but not in the written list.

The examples presented in this section provide a limited level of evidence to suggest that the Chinese students make use of certain vague and informal chunks in their assignments, in line with the learner corpus literature (e.g. Lee and Chen, 2009; Paquot, 2010). From the examples reported here, it seems that for the Chinese students, and to a lesser extent the English students, an awareness of the appropriacy of chunks within different genres of writing is still developing.

Connectors

The term ‘connectors’ is used here to refer to lexical items which have a broadly textual function in connecting parts of the writing (termed ‘linking adverbials’ in Biber et al., 1999: 875). While some linking chunks were noted earlier as idiosyncratic to the case study students (*in one word, that is why*), the data also contains connectors which are salient on reading all four students’ writing due to their relatively high occurrence and which were subsequently found to be used across Chi123; for example:

(14) This can create a positive image for Scotland; *on the other hand*, by referring to the previous experiences. (Mei-Xie)

(15) …in order to create a centre of attention to the tourists. *As a consequence*, it can attract many travelers visiting Liverpool (Hong)

(16) …*On the contrary*, the predominance of SMEs largely carry out on an informal. (Mei-Xie)

Corpus searches revealed these three connectors to be prevalent across Chi123 in comparison with Eng123, and to occur across most disciplines (Figure 5).

*On the other hand* has been discussed in studies of L2 English student writing as a particularly highly-used sequence (e.g. Milton, 1999). This chunk is the most frequent connector in Chi123 (56 occurrences), and is widely dispersed across texts, individuals and
disciplines. For Chinese students, the 4-gram on the other hand may be frequently used as it is often viewed as a translation equivalent to a Mandarin expression meaning ‘two sides of a coin’.

The literature on NNS writing suggests that NNSs generally, and Chinese students in particular, favour particular connectors and that they use these repeatedly (e.g. Gilquin, 2010; Hyland, 2008; Lee and Chen, 2009; Milton, 1999), particularly in sentence-initial position (Milton, 1999). In English language textbooks in China, lists of connectors together with translation equivalents are often provided without information as to the different registers they may be used in (see Leedham and Cai, under review). Since this lack of register differentiation also occurs in the model texts provided by examination boards, reproduced in exam preparation textbooks and subsequently memorized by secondary school students, it is unsurprising that a similar lack of distinction occurs at undergraduate level by Chinese students.

Data references
Both the case study students and students overall in Chi123 used the
same formulaic sequences multiple times to refer the reader to tables, appendices or figures, e.g. as illustrated in table 2, as shown in table (Wei x 2, Mei-Xie x 2), according to (Wei x 4). Figure 6 illustrates this final example, showing that common first and second right collocates for according to are equation, table or similar. The sequence according to occurs significantly more frequently in Chi123 than in Eng123 ($p=.0001$; raw counts of 141 and 242 respectively).

Figure 6: according to in Chi123

The prevalence of formulaic sequences referring to tables, equations or other visual features suggests that the L1 Chinese students make greater use of these elements in their assignments than the L1 English students; this finding is confirmed in research reported in Leedham (2012).

4.2 Findings from keyword analysis

In this second procedure, keywords from the four Chinese students' writing were first extracted by comparing each student’s texts with those in the equivalent discipline corpus of L1 English students’ writing. The resulting four lists of keywords are given in Appendix One. Examining the lists of keywords within the wider co-text of sentence and paragraph, and the context of student assignment-writing gave rise to a number of themes, some of which overlap with the groupings given in 4.1.
**Localized n-grams**

This category includes examples considered to be *idiosyncratic* since they are specific to one of the four case study students, as well as *topic-specific* n-grams occurring in one assignment and *discipline-specific* n-grams occurring within a single discipline. Often, it is hard to distinguish between these subcategories; for example, Mei-Xie’s keywords in Figure 7 occur only within her writing within a single text in HLTM.

Reading the original assignment reveals that the eight concordance lines in Figure 7 occur at the ends of each of eight sections within a single Business assignment. Long chunks of this kind were also apparent in Eng123 within single assignments as students repeat similar information multiple times; in one case the entire abstract and conclusion were identical.

In Wei’s (Engineering) list of keywords, several key chunks are part of longer metalanguage statements; e.g. *aim of the, of the assignment is to design, to develop an understanding of;* all of these chunks occur in assignment introductions in the following pattern:

```
    (the)       aim         of the assignment      is to design
       object     is to develop an understanding of
```

These chunks appear to be Wei’s preferred way of setting out the aim of an assignment. While they occur in other texts within Chi123 and Eng123, the n-grams are key in Wei’s writing when compared to the larger corpus of English-Engineering texts.

More topic-specific n-grams are those occurring in a particular
subject-area within a discipline, and usually within single texts. For example, in Mei-Xie’s HLTM writing, the chunk *IHG annual report* is concerns a company report, and occurs five times within an assignment entitled ‘Executive Summary: InterContinental Hotels Group Plc (IHG)’. Similarly, many of Hong’s n-grams are topic-specific and found in single texts e.g. *Marriott Liverpool city centre* (x 17) and *the Liverpool tourism industry* (x 6). All four of Feng’s keywords are topic-specific, with three occurring in a single text. In fact, the absence of non-localized keywords in the list for Feng suggests there is little difference between her writing and that of the reference corpus in terms of the shared ‘aboutness’ of the writing.

The two HLTM students, Hong and Mei-Xie, use n-grams relating to the whole discipline or vocational area more than the L1 English HLTM corpus; for example, *the tourism industry* (Hong), *the hospitality industry* (Mei-Xie), *recruitment and selection* (Mei-Xie) and *in the hospitality industry* (Mei-Xie). It could be the case that these two students make greater reference to the whole area of hospitality management, or perhaps in English-HLTM a wider range of n-grams is used to discuss the whole discipline, though this was not apparent from the keyword analysis. Studies of lexical chunks extracted from different disciplines provide useful comparisons here (e.g. Simpson-Vlach and Ellis, 2010; Cortes, 2004) though little has been done in the Hospitality area.

**Connectors**

In contrast to the multiple connectors highlighted in method one, the only keyword with a primary connecting function to be revealed through keyword analysis is *on the other hand*. This chunk is key in Mei-Xie’s writing and, while present in the other three students’ texts, is not a keyword.

**Data references**

The keyword lists for Wei and Mei-Xie each include directives to data given in assignment appendices (e.g. *in the appendix, with reference to appendix*). While Wei’s chunks are spread throughout the ten assignments, most of Mei-Xie’s occur in a year 2 text and are part of
directives guiding the reader to multiple appendices (the same proposal text as discussed under localized n-grams above). Since many students did not include appendices with their BAWE submissions, it is not possible to calculate whether Chinese students are more likely to use multiple and/or longer appendices, or whether they simply reference these more frequently using particular chunks.

Wei’s keywords also include references to equations (or eq) and tables (e.g. were recorded as below, was calculated with eq) and several keywords contain a formula\(^5\). A keyword search in Chinese-Engineering reveals that references to visual features are key to all L1 Chinese Engineering students, suggesting that these are used or at least referred to more prevalently than in English-Engineering (see also Leedham, 2012).

**Passives**

Two of the keyword lists contain some passive statements, e.g. be worked out (Wei), can be calculated (Wei) and it is believed that (Mei-Xie). Here, the latter was investigated further in her writing using the WordSmith concordancer to search for the string it is * that with the asterisked item limited to verbs (Figure 8).

---

\(^5\) Note that all mathematical formulae are replaced in BAWE by the capitalized FORMULA.
The same search in English-HLTM resulted in eight chunks, equating to just one seventh of Mei-Xie’s use of *it is *that* after normalization. Anticipatory *it* clauses seem to be Mei-Xie’s preferred way of expressing her views, perhaps since these are less overt than employing personal pronouns (Hewings and Hewings, 2002; see also Groom, 2006; Römer, 2009). An additional reason for Chinese students’ avoidance of the individual voice presented through *I* and a preference for the collective *we* is the influence of a collectivist culture in which the individual view is subsumed within the group (e.g. Snively 1999).

5. Conclusions
The two methods for identifying lexical chunks in the case study students’ writing uncover some common categories. Both reading the texts for salient chunks and using keyword searches suggest that these students, and in some cases Chinese students more generally, employ particular connectors (though only *on the other hand* is a keyword), and make greater reference to data contained in appendices, tables, or figures. Idiosyncratic chunks such as *in one word* and *catch the world’s eyes* were found through the intuitive reading of the first method as these sequences are infrequently-used, yet may have a disproportionate impact on the reader’s view of the writing. Close reading of the texts additionally suggests that the Chinese students use some vague and informal chunks (e.g. *more or less*), though the data here is limited. Items occurring sufficiently frequently in a single student’s writing to be extracted as keywords were usually topic-specific (e.g. *IHG annual report*); the extraction of keywords across the four students’ writing highlights repeated chunks across texts which may be useful for pedagogic purposes (e.g. *the aim of the assignment is to design*).

Both methods for identifying marked lexical chunks provide starting points in exploring features of the four students’ texts, all of which have been judged by discipline specialists to be proficient undergraduate assignments. Notably, each method benefits from the additional checks provided by the other: salient formulaic sequences can be searched for using corpora to confirm the extent of use, while keywords benefit from exploration within the context of whole texts. Viewing texts as complete Word documents gives a sense of the whole assignment as it was read by the discipline lecturer, and highlights
features such as tables, chart and lists since these are visually different from continuous running prose. In this sense a corpus investigation is reductive since multimodal features such as the layout of text and visuals on the page are downplayed or lost.

Reading the assignments to intuitively select formulaic sequences was difficult in unfamiliar disciplines; in such cases the analyst could make use of subject specialists and a reference corpus or academic formulas list (e.g. Simpson-Vlach and Ellis, 2010). For example, Wei’s Engineering writing was difficult for the non-Engineer to determine whether specialized terms are discipline-specific sequences or whether they have been coined by one student (and are perhaps formulaic sequences for that student). Appendix Two shows an attempt to categorize contiguous formulaic sequences in a 250-word introduction. This difficulty in recognising sequences has pedagogical implications since writing tutors seldom have the same disciplinary background as their students. While it is likely that language users within a discourse community such as Engineering academics agree on a large number of shared core sequences there are also many peripheral sequences which are particular to subsets or to individuals within the group. It is unsurprising, then, that individuals often identify different sequences and set sequence boundaries differently (e.g. Foster, 2001; Leedham, 2006) since each individual experiences different language ‘primings’ according to their previous linguistic exposure (Hoey, 2005).

In contrast, beginning with a keyword search is quick, easily replicable and does not rely on discipline-specific knowledge from the analyst. However, subjective choices must still be made: the linguist must select or compile a representative corpus and perhaps a reference corpus, choose software and set parameters within the software, as well as limiting the searches to a manageable amount of data. While corpus analysts have always explained their data using intuition (Borsley and Ingham, 2002), the corpus itself is rarely read and the cohesion of individual texts is lost. Whereas all concordance lines are treated equally, when reading an assignment a single, marked chunk may have a disproportionate impact on the reader.

One fruitful direction for individuals is the exploration of a corpus of their own writing. For example, the use of passive constructions (e.g. *it is believed that*) points to a potential difference in the expression of stance in Mei-Xie’s writing when compared to the reference corpus. The
use of data-driven learning is explored in Lee and Swales (2006) in their
description of a course entitled ‘exploring your own discourse world’ in
which students compiled corpora of their own writing and compared this
to reference corpora of research articles in their discipline. Similarly,
Coniam (2004) built a corpus of his own writing, describing the process
as ‘technology-enhanced rhetorical consciousness-raising’ (p.72). While
writing or discipline tutors are unlikely to have the time to check their
intuitive reading in a corpus of student writing, classes featuring data-
driven learning can enhance student recognition of their own writing
style.

Recursivity of method, such as corpus searches followed by
reading and more corpus searches has been described by Matthiessen
(2006: 110) as a ‘two-pronged approach’ and combines some of the
benefits of each method. Knowing exactly what is in the corpus, in what
proportions, and being able to read whole texts is important in providing
insights for further corpus exploration, and at the very least, reminds the
user that they are looking at real language taken out of its original
context. While the small-scale nature of this study enabled the
assignments to be individually read, the benefits of this method can be
applied to larger corpora by reading a selection of the texts in order to
complement corpus analysis. This paper argues that a multi-method
approach allows more to be discovered and justified, as illustrated by
Hunston’s comment that corpora ‘are invaluable for doing what they do,
and what they do not do must be done in another way’ (2002: 20).

Note: The British Academic Written English (BAWE) corpus is a
collaboration between the universities of Warwick, Reading and Oxford
Brookes. It was collected as part of the project, ‘An Investigation of
Genres of Assessed Writing in British Higher Education’ funded by the

Acknowledgements
I am grateful for the suggestions made by Lina Adinolfi, Jean Hudson
and two anonymous NJES reviewers.
References


Nesselhauf, Nadja. (2005), Collocations in a Learner Corpus (Studies in Corpus Linguistics) (John Benjamins).


Appendix One: Keywords in the 4 students’ texts

*Wei: Engineering*

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Wei Freq</th>
<th>Wei Texts</th>
<th>L1Eng Freq</th>
<th>L1Eng Texts</th>
<th>Keyness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in the appendix</td>
<td>14</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td>the one with</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>FORMULA FORMULA FORMULA FORMULA</td>
<td>47</td>
<td>4</td>
<td>197</td>
<td>25</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>was calculated with eq.</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>is shown as</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>the other one</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>briefing sheet in appendix</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>in steps of</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>aim of the</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>than the one with</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>to develop an understanding of</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td>can be calculated respectively</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>of the assignment is to design</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>in this design, the</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>15</td>
<td>could be worked out</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>16</td>
<td>tables of data</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>17</td>
<td>were recorded as below</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>18</td>
<td>as below FORMULA</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>FORMULA FORMULA FORMULA FORMULA</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>20</td>
<td>applying equation</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>21</td>
<td>the change of</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>22</td>
<td>therefore, the bending</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>23</td>
<td>of these two</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>24</td>
<td>has to be</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>25</td>
<td>be worked out</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>26</td>
<td>in this laboratory</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>20</td>
</tr>
</tbody>
</table>
### Frequency of Texts

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Ping Freq.</th>
<th>Ping Texts</th>
<th>L1Eng HLTM Freq.</th>
<th>L1Eng HLTM Texts</th>
<th>Keyness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the hospitality industry</td>
<td>16</td>
<td>3</td>
<td>42</td>
<td>12</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>recruitment and selection</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>in the hospitality industry</td>
<td>10</td>
<td>2</td>
<td>20</td>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>please see appendix</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>with reference to appendix</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>higher than the original figure of</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>the new level of net profit</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>quality of service</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>the cost of</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>to the guests</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>it is believed that</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>of the employees</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>13</td>
<td>there will be</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>14</td>
<td>of the group</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>15</td>
<td>to reach the break even point</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td>on the other hand</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>17</td>
<td>will be a</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>18</td>
<td>high quality of service</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>cost of sales</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>the nature of</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>21</td>
<td>Watson and Head</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>22</td>
<td>IHG annual report</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>23</td>
<td>a higher contribution</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>24</td>
<td>Atrill and McLaney</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>P E ratio</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>served to the</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
</tbody>
</table>
### Hong: HLTM

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Hong Freq.</th>
<th>Hong Texts</th>
<th>L1Eng Freq.</th>
<th>L1Eng Texts</th>
<th>Keyness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Liverpool city centre</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>Marriott Liverpool city centre</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>city centre hotel</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Liverpool city centre hotel</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>Marriott Liverpool city centre hotel</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>6</td>
<td>Oxford and Bath</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>7</td>
<td>European capital of</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>North East Somerset</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>European capital of culture</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>Burgess and Bryant</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>Dunn and Brooks</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>12</td>
<td>Liverpool tourism industry</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>13</td>
<td>night stays arriving</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>14</td>
<td>North East Somerset council</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>the European capital of</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>16</td>
<td>the Liverpool tourism industry</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>17</td>
<td>in the city centre</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>18</td>
<td>the city centre</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>park and ride</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>20</td>
<td>in terms of the</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>in the Liverpool tourism industry</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>and Bath are</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>23</td>
<td>bargaining power of</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>24</td>
<td>city centre is</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>25</td>
<td>the tourism industry</td>
<td>13</td>
<td>3</td>
<td>16</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

### Feng: Food Science

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Feng Freq.</th>
<th>Feng Texts</th>
<th>L1Eng Freq.</th>
<th>L1Eng Texts</th>
<th>Keyness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>of coliform bacteria</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>Wang et al.</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>the recommended RNI</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>the air bubbles</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
</tbody>
</table>
Appendix Two: Chunked paragraphs from Wei’s writing

Note: The emboldened words indicate formulaic sequences.

Introduction

A design methodology for a gearbox is presented in this report. The input horse power, the input speed and net reductions in the gearbox are the parameters to be specified. A gearbox takes an input shaft rotating and converts it via a gear train into up to three outputs, the process of designing a gearbox is to figure out which ratios are needed and to implement those ratios in the form of positioning various sizes of connected gears. The specification of the gearbox depends on its area of application.

In this report, a gearbox is designed for a commercial meat slicer which has its final shaft rotating at between 80 and 100 rev/min. The input of the meat slicer is a constant speed AC motor running at 1800 rev/min and delivering 1.2 kW. A few points have to be considered on this system, the size of the gearbox is severe restricted, since it has to go onto a work surface where there is severe competition for space. And the motor may be in-line or at right angles to the grinder. Furthermore, the duty is expected to be up to 6 hours per day.

In this design, firstly, the gear ration was decided, and a specimen manual calculation was taken to check bending and surface stress, the result was compared with Gears Program. After that, according to the calculated gear teeth loads, the design of shaft and bearings were discussed. Finally, the designed gearbox was drawn in Solidworks.