Searching corpora of Chinese and British writers for lexicalised language
中英学生写作语料库中惯用语使用的探析

Maria Leedham
Outline

- Study overview and RQs
- Lexicalised language
- Comparing the writing
- Questionnaire data
The Study

My Background:

• Research Assistant on a corpus project collecting student assignments and interviewing lecturers, (‘British Academic Written English’ project)

• English for Academic Purposes teacher in universities teaching Chinese students.

• My PhD looks at Chinese and British students’ assessed writing in U.K. universities.

• Years 1,2,3 & PG; narrow to 3 discipline areas; corpus study plus questionnaire and interviews.
Research Questions

• 1. In what ways do British and Chinese undergraduate and Master’s students in U.K. universities differ in their use of lexicalised language in the academic writing of three selected disciplines?

• 2. How do these students develop their use of lexicalised language from year 1 to year 3 of undergraduate study?

• 3. What are the pedagogical implications for teachers of academic writing?
Building the Corpus 1

• A corpus is a balanced collection of texts, usually stored in electronic form.

• My two corpora will comprise 300 assignments from Chinese students and 300 from British students – around 1 million words per corpus.

• Assignments will be matched for discipline, year groups, age and gender of students. All assignments will be anonymised.

• The two corpora can be searched using corpus linguistics software e.g. WordSmith Tools v.5
Building the Corpus 2

• British Academic Written English corpus (Bawe) as a starting point :-
  - 4 Universities (Warwick, Reading, Oxford Brookes & Coventry)
  - 32 different disciplines
  - 3000 assignments
  - II:i or I level.

• Currently – collecting more assignments (70 so far) through contacts, CSSA, FaceBook.
## Overview of Chinese corpus

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Master’s</th>
<th>Total per discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Life Sciences</strong> (food, biological &amp; plant sciences)</td>
<td>21</td>
<td>17</td>
<td>15</td>
<td>17</td>
<td>70</td>
</tr>
<tr>
<td><strong>Social Sciences</strong> (business, economics and hospitality)</td>
<td>19</td>
<td>11</td>
<td>52</td>
<td>25</td>
<td>107</td>
</tr>
<tr>
<td><strong>Physical Sciences</strong> (cybernetics, all engineering, computing)</td>
<td>12</td>
<td>16</td>
<td>48</td>
<td>5</td>
<td>81</td>
</tr>
<tr>
<td><strong>Arts &amp; Humanities</strong> (linguistics, ICT in education, theatre studies)</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td><strong>Totals per year group</strong></td>
<td>52</td>
<td>44</td>
<td>120</td>
<td>63</td>
<td><strong>Total: 279</strong></td>
</tr>
</tbody>
</table>

Maria Leedham
Outline

• Study overview and RQs

• Lexicalised language

• Comparing the writing

• Questionnaire data
Lexicalised language as ‘chunks’

- Psychology research suggests humans group phenomena together, i.e. we process information in "chunks" or meaningful units of information (Miller, 1956).
  e.g. chess (remembering sequences of moves as chunks), telephone numbers, stretches of music

- **Thus** - “it is possible to expand the total amount of information by packing more and more information into one chunk". (Howard 1983:104).
  e.g. A chess expert has sequences of longer moves than the novice player; a skilled language user chooses from a variety of long lexicalized sentence stems while a lower-level user struggles to communicate with a few short chunks.

- **So** - we use “an abundant resource (memory to store prefabricated chunks of language) to compensate for a limited one (processing capacity)". (Schmitt and McCarthy, 1997:230).
Examples of lexicalised language or ‘chunks’

- it is interesting to note
- in other words
- in order to
- and so on
- on the other hand
- the fact is
- it has been argued that
- based on the
- human beings are
- is there a link between
- such as
- individual needs
- are there
- out of
How much language is lexicalised?

4-5%  Moon , 1998  “FEI”= fixed expression and idiom

“holistic units of two or more words”

Includes frozen collocations, proverbs, routine formulae, sayings, similes

e.g. armed to the teeth, foot the bill, red herring, at home, little by little,

Excludes compound nouns, adjectives, verbs; phrasal verbs, foreign phrases, multi-word inflectional forms e.g. civil servant, self-raising, had been lying, more careful

80%  Altenberg, 2001  “recurrent word combinations”

Includes “any continuous string of words occurring more than once in identical form”

e.g. I think that, do you know, out of the, what sort of, because I mean, and then I,

Excludes any word strings occurring once only.
How do we find lexicalised language?

- ScriptLog software
- Eye-tracking

- Structural features
  - e.g. fixedness, non-compositionality

- Phonological features,
  - e.g. speech rate, clarity of articulation, intonation contour, lack of internal pausing

- Frequency counts

- Human intuition
Lexicalised Language

From frequency counts

- ‘Clusters’ are “words which are found repeatedly together in each others' company, in sequence”.

- Frequently occur across structural groups
  e.g. way we speak is there a

- Evidence to suggest that not all of these clusters are stored as wholes in the mental lexicon.

From human intuition

- A ‘formulaic sequence’ is “a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar”

- Sequences do not cross structural boundaries - e.g. the way we speak

- Sequences are “psychologically real” and are thought to be stored as wholes in the mental lexicon.
Lexicalisation = “the process by which a string of words and morphemes becomes institutionalized as part of the language and develops its own specialist meaning or function.” Moon, 1998:36.
Functions of lexicalised language

“The more novel our output is for the hearer, the more likely it is to be misunderstood” (Wray, 2002:94).

Maria Leedham
Outline

• Study overview and RQs
• Lexicalised language
• Comparing the writing
• Questionnaire data
<table>
<thead>
<tr>
<th>No.</th>
<th>Chunk</th>
<th>Freq.</th>
<th>Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>on the other hand</td>
<td>61</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>as a result of</td>
<td>26</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>as well as the</td>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>at the same time</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>one of the most</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>it is important to</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>as one of the</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>at the end of</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>it is necessary to</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>can be used to</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>it is difficult to</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>can be seen that</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>it can be seen</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>it is believed that</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>this is due to</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Chunk</th>
<th>Freq.</th>
<th>Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>at the end of</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>as a result of</td>
<td>25</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>it is important to</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>at the same time</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>can be used to</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>it can be seen</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>this is due to</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>can be seen that</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>it is possible to</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>on the other hand</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>it is clear that</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>may be due to</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>
It is vital for motivation theories to consider four influential individual needs in the workplace, which are the competence, achievement, affiliation, and money motives. ‘The competence motive’ is the desire for job mastery and professional growth. Robert White suggests the competence motive to be based on the assumption that a person is not only “a vehicle for a set of instincts” (Gellerman 1963: 111), but is also eager on discovering and fulfilling their potential. It is assumed that humans are keen on manipulating their environment to pursue goals. Thus, competence is a key motive affecting job success, because people who have faith in their own ability to influence the environment do tend to succeed.

On the other hand, individuals with a strong achievement motive perceive accomplishment as an end. Achievement-motivated employees search for the opportunities to obtain successes that are “hard but are not unobtainable” (Gullerman 1965: 126), and thus, tend to outperform others by constantly challenging themselves. The reasonable degree of risk involved in the goal-attainment process encourages employees to set realistic goals and to maximize their abilities.

Affiliation is another individual need, which refers to the “social drive to be associated with others in interdependent relationships, involving using others for help or support without making them responsible for problems” (MerckSource, 2006). Affiliation can be considered as a means to an end or an ends itself – people socialize with fellow workers for specific purposes, such as favors or protection, or simply for enjoyment (Gellerman, 1965).
Concordance lines for ‘on the other hand’

**ChiCor**

<table>
<thead>
<tr>
<th>No.</th>
<th>Concordance</th>
<th>Search term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>increased fat deposition with age.</td>
<td>On the other hand, males tend to have</td>
</tr>
<tr>
<td>2</td>
<td>and reduces the cost dramatically.</td>
<td>On the other hand, Ford’s product still</td>
</tr>
<tr>
<td>3</td>
<td>y change when the product is used.</td>
<td>On the other hand, it’s really difficult</td>
</tr>
<tr>
<td>4</td>
<td>cell’s viewing angle performance.</td>
<td>On the other hand, the viewing</td>
</tr>
</tbody>
</table>

**EngCor**

<table>
<thead>
<tr>
<th>No.</th>
<th>Concordance</th>
<th>Search term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>by (Williams and Feltmate, 1992).</td>
<td>On the other hand, if there is a form of</td>
</tr>
<tr>
<td>2</td>
<td>ates smoking tobacco and cancer.</td>
<td>On the other hand, the drug testing for</td>
</tr>
<tr>
<td>3</td>
<td>n in Fig 14 and 15) formica fusca,</td>
<td>on the other hand, was found mainly in</td>
</tr>
<tr>
<td>4</td>
<td>very much an altruistic view. Then</td>
<td>on the other hand, is it ethical to stop</td>
</tr>
<tr>
<td>N</td>
<td>Word</td>
<td>Freq.</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>it can be seen that the</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>it is important to note that</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>set of tasks questions exercises non</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>gender on the way we speak</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>is there a link between them</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>speak is there a link between</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>the social variables of class and</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>the way we speak is there</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>way we speak is there a</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>we speak is there a link</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>and gender on the way we</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>class and gender on the way</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>dot plot to show the mean</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>is due to the fact that</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>of class and gender on the</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>of the social variables of class</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>signs but do they have language</td>
<td>10</td>
</tr>
<tr>
<td>18</td>
<td>variables of class and gender on</td>
<td>10</td>
</tr>
<tr>
<td>19</td>
<td>in order to be able to</td>
<td>9</td>
</tr>
</tbody>
</table>
What is the effect of the social variables of gender and class on the way we speak? Is there a link between them?

The effects of gender and class on the way we speak is a question that has engaged much time with linguists and

As well as social class, Gender also obviously has a huge effect on the way we speak - both in single sex and mixed sex

way we speak as well as helping us link the effects of social class and gender on the way we speak. Lakoff has claimed

What is the effect of the social variables of class and gender on the way we speak? Is there a link between them?

I will initially discuss the effect of social class on the way we speak and followed by gender as the other social variable.
Beyond clusters?

- Clusters, n-grams, = contiguous words
- ConcGrams, Greaves & Warren, 2007, e.g. A**B, B*A,

**PoS-gram**
e.g. prep.+det.+noun+of
at the end of, as a result of,

**Semantic sequences**
e.g. time + journey time + transport
In Summer it’s a half-hour journey by bus
Outline

• Study overview and RQs

• Lexicalised language

• Comparing the writing

• Questionnaire data
Questionnaire Data

• 170 questionnaire responses so far to online survey of British and Chinese students’ view on writing assignments.

• Over 40 universities and over 20 disciplines are represented.

• Where and how were you taught academic writing?
• At secondary school...We were taught vocabulary to use. (year 2, Cantonese speaker)
• I have never been taught about essay writing ... I still don't even know if I am doing it right but as I have passed so far I guess it’s ok. (year 3, native English speaker)
Have you changed your way of planning and writing your assignments?

- I had been used to writing in Chinese first and then translating. But afterward I switched to writing and planning at the same time... *(Mandarin)*

- Yes - I plan them before I write. Also, I've absorbed some of the 'style' appropriate to the discipline. *(NES)*

- I plan, write a draft, leave it, then go back and re-read and adjust it. repeat as necessary. *(NES)*
Generally, how do you feel about assignment writing at the moment?

- I enjoy it, but I think I spend more time on it than I should. *(year 2, NES)*
- After I finish it, I have a sense of accomplishment. *(Mandarin)*
- I really like writing assignments because you feel a sense of satisfaction when they are done especially if you get a good mark afterwards. *(NES)*
- I like assignment writing, I prefer it to exams. *(NES)*
- I do not enjoy it because I always have to rush to meet the deadline. *(Cantonese)*
- I would have enjoyed if there were not so many to do at one time. *(Cantonese)*
- I've always hated English as I am not strong at it. I still feel the same after doing university assignments but at least they are in a subject I am interested in. *(NES)*
Potential outcomes

- Help UK university lecturers to understand differences between Chinese and British ways of writing.

- Increase knowledge on common chunks in British and Chinese students’ writing.

- Assist EAP teachers by identifying features of successful third year student writing in both British and Chinese students’ writing.

- Aid materials writers in designing coursebooks.
Issues to consider

• **Possible case studies:**
  - consider ‘missed chunks’ in students’ writing
  - ask students to identify lexicalised language

• **How typical are students who submitted assignments?**
  - What about the less proficient ones from either language group?
  - How homogenous is this group of ‘Chinese students’?

• **What software?**
  Use WordSmith Tools to identify clusters + additional software to identify concgrams & semantic sequences

• **What am I looking for?**
  Lexicalised language is frequently-used language in different types of writing e.g. year 1 and year 3 of study,
  Different disciplines, Chinese and British students.
  Classify chunks according to function – useful for teaching?
References

• British Academic Written English project. http://www2.warwick.ac.uk/fac/soc/celte/research/bawe
• Leedham, M.E. 2006 ‘Do I Speak Better?’: A longitudinal study of lexical chunking in the spoken language of two Japanese students’. In The East Asian Learner. 2(2) http://www.brookes.ac.uk/schools/education/eal/eal-2-2/
• Surveymonkey. https://www.surveymonkey.com
Acknowledgement

• The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800)