Corpora in Foreign Language Teaching

- Vivienne Rogers
  - School of Modern Languages, Newcastle University
  - www.viviennerogers.info
  - vivienne.rogers@education.ox.ac.uk

- Zoe Handley
  - Department of Education, University of Oxford
  - zoe.handley@education.ox.ac.uk
  - http://humbox.ox.ac.uk/profile/291

Linguistics Association of Great Britain Conference 2010

Plan

- What is a corpus?
- Basic corpus techniques
- Corpora in language learning
- Data-driven language learning (DDL)
- What the research says about DDL?
- Benefits of DDL
- Limitations of DDL
- Working within the limitations of DDL

Corpus linguistic techniques

- Concordancing
  - “using corpus software to find every occurrence of a particular word or phrase” (O’Keefe et al., 2001: 8)

- Word frequency counts and word lists

- Key word analysis
  - “Key words ... are those whose frequency is unusually high in comparison with some norm” (O’Keefe et al., 2001: 12)

- Cluster analysis
  - Cluster analysis allows the user to generate a list of the most frequent
  3-, 4-, 5-, or 6-word combinations (n-grams, word/lexical clusters/bundles) from a corpus, i.e. collocations and colligations (O’Keefe et al., 2001)


Vivienne Rogers & Zoe Handley

“Only when words are in their habitual environments, presented in their most frequent forms and their relational patterns and structures, can they be learnt effectively, interpreted properly and used appropriately”

(Wu, 1992: 32)

What is a corpus?

“any collection of more than one text can be called a corpus: the term ‘corpus’ is simply the Latin for ‘body’, hence a corpus may be defined as any body of text”

(McEnery and Wilson, 2001: 29)

“... a collection of pieces of language, selected and ordered according to explicit linguistic criteria in order to be used as a sample of language”

(Sinclair, 1996)

• Reference corpus
  - British National Corpus, Brown Corpus
  - Balanced sample, machine-readable form, annotated
Corpus linguistic techniques

- Concgamming
  - A "congram" is all of the permutations of constituency variation and positional variation generated by the association of two or more words" (Graves and Warren, 2007: 296)
- Lexico-grammatical profiles
  - Collocates
  - Chunks/idioms
  - Syntactic restrictions
  - Semantic restrictions
  - Semantic prosody

Concordancing

"using corpus software to find every occurrence of a particular word or phrase" ... "The search word of phrase is often referred to as the 'node' and concordance lines are usually presented with the word/phrase in the centre of the line with seven or eight words presented at either side. These are known as Key-Word-In-Context displays (or KWIC concordances"

(O’Keefe et al., 2001: 8)

Key Word Analysis

"Key words ... are those whose frequency is unusually high in comparison with some norm"

(O’Keefe et al., 2001: 12)

- Wordsmith Tools (Scott, 1999)
  - Compares the word list obtained from a small corpus with that obtained from a large reference corpus
  - Applications: genre analysis, forensic linguistics, stylistics, content analysis, text retrieval, and Languages for Specific Purposes

Key words from economics lecture relative to corpus of academic lectures


<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tax</td>
<td>8</td>
<td>poor</td>
</tr>
<tr>
<td>2</td>
<td>Income</td>
<td>9</td>
<td>thousand</td>
</tr>
<tr>
<td>3</td>
<td>Average</td>
<td>11</td>
<td>equity</td>
</tr>
<tr>
<td>4</td>
<td>System(s)</td>
<td>10</td>
<td>impact</td>
</tr>
<tr>
<td>5</td>
<td>Basic</td>
<td>12</td>
<td>under</td>
</tr>
<tr>
<td>6</td>
<td>Rate</td>
<td>13</td>
<td>both</td>
</tr>
<tr>
<td>7</td>
<td>Supply</td>
<td>14</td>
<td>figures</td>
</tr>
<tr>
<td>8</td>
<td>Higher</td>
<td>15</td>
<td>Higher</td>
</tr>
<tr>
<td>9</td>
<td>Higher</td>
<td>16</td>
<td>Higher</td>
</tr>
<tr>
<td>10</td>
<td>Correlation</td>
<td>19</td>
<td>Correlation</td>
</tr>
<tr>
<td>11</td>
<td>Correlation</td>
<td>20</td>
<td>Correlation</td>
</tr>
<tr>
<td>12</td>
<td>Correlation</td>
<td>21</td>
<td>Correlation</td>
</tr>
<tr>
<td>13</td>
<td>Correlation</td>
<td>22</td>
<td>Correlation</td>
</tr>
<tr>
<td>14</td>
<td>Correlation</td>
<td>23</td>
<td>Correlation</td>
</tr>
<tr>
<td>15</td>
<td>Correlation</td>
<td>24</td>
<td>Correlation</td>
</tr>
<tr>
<td>16</td>
<td>Correlation</td>
<td>25</td>
<td>Correlation</td>
</tr>
<tr>
<td>17</td>
<td>Correlation</td>
<td>26</td>
<td>Correlation</td>
</tr>
<tr>
<td>18</td>
<td>Correlation</td>
<td>27</td>
<td>Correlation</td>
</tr>
<tr>
<td>19</td>
<td>Correlation</td>
<td>28</td>
<td>Correlation</td>
</tr>
<tr>
<td>20</td>
<td>Correlation</td>
<td>29</td>
<td>Correlation</td>
</tr>
<tr>
<td>21</td>
<td>Correlation</td>
<td>30</td>
<td>Correlation</td>
</tr>
<tr>
<td>22</td>
<td>Correlation</td>
<td>31</td>
<td>Correlation</td>
</tr>
<tr>
<td>23</td>
<td>Correlation</td>
<td>32</td>
<td>Correlation</td>
</tr>
<tr>
<td>24</td>
<td>Correlation</td>
<td>33</td>
<td>Correlation</td>
</tr>
<tr>
<td>25</td>
<td>Correlation</td>
<td>34</td>
<td>Correlation</td>
</tr>
<tr>
<td>26</td>
<td>Correlation</td>
<td>35</td>
<td>Correlation</td>
</tr>
<tr>
<td>27</td>
<td>Correlation</td>
<td>36</td>
<td>Correlation</td>
</tr>
<tr>
<td>28</td>
<td>Correlation</td>
<td>37</td>
<td>Correlation</td>
</tr>
<tr>
<td>29</td>
<td>Correlation</td>
<td>38</td>
<td>Correlation</td>
</tr>
<tr>
<td>30</td>
<td>Correlation</td>
<td>39</td>
<td>Correlation</td>
</tr>
<tr>
<td>31</td>
<td>Correlation</td>
<td>40</td>
<td>Correlation</td>
</tr>
<tr>
<td>32</td>
<td>Correlation</td>
<td>41</td>
<td>Correlation</td>
</tr>
<tr>
<td>33</td>
<td>Correlation</td>
<td>42</td>
<td>Correlation</td>
</tr>
<tr>
<td>34</td>
<td>Correlation</td>
<td>43</td>
<td>Correlation</td>
</tr>
<tr>
<td>35</td>
<td>Correlation</td>
<td>44</td>
<td>Correlation</td>
</tr>
<tr>
<td>36</td>
<td>Correlation</td>
<td>45</td>
<td>Correlation</td>
</tr>
<tr>
<td>37</td>
<td>Correlation</td>
<td>46</td>
<td>Correlation</td>
</tr>
<tr>
<td>38</td>
<td>Correlation</td>
<td>47</td>
<td>Correlation</td>
</tr>
<tr>
<td>39</td>
<td>Correlation</td>
<td>48</td>
<td>Correlation</td>
</tr>
<tr>
<td>40</td>
<td>Correlation</td>
<td>49</td>
<td>Correlation</td>
</tr>
<tr>
<td>41</td>
<td>Correlation</td>
<td>50</td>
<td>Correlation</td>
</tr>
<tr>
<td>42</td>
<td>Correlation</td>
<td>51</td>
<td>Correlation</td>
</tr>
<tr>
<td>43</td>
<td>Correlation</td>
<td>52</td>
<td>Correlation</td>
</tr>
<tr>
<td>44</td>
<td>Correlation</td>
<td>53</td>
<td>Correlation</td>
</tr>
<tr>
<td>45</td>
<td>Correlation</td>
<td>54</td>
<td>Correlation</td>
</tr>
<tr>
<td>46</td>
<td>Correlation</td>
<td>55</td>
<td>Correlation</td>
</tr>
<tr>
<td>47</td>
<td>Correlation</td>
<td>56</td>
<td>Correlation</td>
</tr>
<tr>
<td>48</td>
<td>Correlation</td>
<td>57</td>
<td>Correlation</td>
</tr>
<tr>
<td>49</td>
<td>Correlation</td>
<td>58</td>
<td>Correlation</td>
</tr>
<tr>
<td>50</td>
<td>Correlation</td>
<td>59</td>
<td>Correlation</td>
</tr>
<tr>
<td>51</td>
<td>Correlation</td>
<td>60</td>
<td>Correlation</td>
</tr>
<tr>
<td>52</td>
<td>Correlation</td>
<td>61</td>
<td>Correlation</td>
</tr>
<tr>
<td>53</td>
<td>Correlation</td>
<td>62</td>
<td>Correlation</td>
</tr>
<tr>
<td>54</td>
<td>Correlation</td>
<td>63</td>
<td>Correlation</td>
</tr>
<tr>
<td>55</td>
<td>Correlation</td>
<td>64</td>
<td>Correlation</td>
</tr>
<tr>
<td>56</td>
<td>Correlation</td>
<td>65</td>
<td>Correlation</td>
</tr>
<tr>
<td>57</td>
<td>Correlation</td>
<td>66</td>
<td>Correlation</td>
</tr>
<tr>
<td>58</td>
<td>Correlation</td>
<td>67</td>
<td>Correlation</td>
</tr>
<tr>
<td>59</td>
<td>Correlation</td>
<td>68</td>
<td>Correlation</td>
</tr>
<tr>
<td>60</td>
<td>Correlation</td>
<td>69</td>
<td>Correlation</td>
</tr>
<tr>
<td>61</td>
<td>Correlation</td>
<td>70</td>
<td>Correlation</td>
</tr>
<tr>
<td>62</td>
<td>Correlation</td>
<td>71</td>
<td>Correlation</td>
</tr>
<tr>
<td>63</td>
<td>Correlation</td>
<td>72</td>
<td>Correlation</td>
</tr>
<tr>
<td>64</td>
<td>Correlation</td>
<td>73</td>
<td>Correlation</td>
</tr>
<tr>
<td>65</td>
<td>Correlation</td>
<td>74</td>
<td>Correlation</td>
</tr>
<tr>
<td>66</td>
<td>Correlation</td>
<td>75</td>
<td>Correlation</td>
</tr>
<tr>
<td>67</td>
<td>Correlation</td>
<td>76</td>
<td>Correlation</td>
</tr>
<tr>
<td>68</td>
<td>Correlation</td>
<td>77</td>
<td>Correlation</td>
</tr>
<tr>
<td>69</td>
<td>Correlation</td>
<td>78</td>
<td>Correlation</td>
</tr>
<tr>
<td>70</td>
<td>Correlation</td>
<td>79</td>
<td>Correlation</td>
</tr>
<tr>
<td>71</td>
<td>Correlation</td>
<td>80</td>
<td>Correlation</td>
</tr>
<tr>
<td>72</td>
<td>Correlation</td>
<td>81</td>
<td>Correlation</td>
</tr>
<tr>
<td>73</td>
<td>Correlation</td>
<td>82</td>
<td>Correlation</td>
</tr>
<tr>
<td>74</td>
<td>Correlation</td>
<td>83</td>
<td>Correlation</td>
</tr>
<tr>
<td>75</td>
<td>Correlation</td>
<td>84</td>
<td>Correlation</td>
</tr>
<tr>
<td>76</td>
<td>Correlation</td>
<td>85</td>
<td>Correlation</td>
</tr>
<tr>
<td>77</td>
<td>Correlation</td>
<td>86</td>
<td>Correlation</td>
</tr>
<tr>
<td>78</td>
<td>Correlation</td>
<td>87</td>
<td>Correlation</td>
</tr>
<tr>
<td>79</td>
<td>Correlation</td>
<td>88</td>
<td>Correlation</td>
</tr>
<tr>
<td>80</td>
<td>Correlation</td>
<td>89</td>
<td>Correlation</td>
</tr>
<tr>
<td>81</td>
<td>Correlation</td>
<td>90</td>
<td>Correlation</td>
</tr>
<tr>
<td>82</td>
<td>Correlation</td>
<td>91</td>
<td>Correlation</td>
</tr>
<tr>
<td>83</td>
<td>Correlation</td>
<td>92</td>
<td>Correlation</td>
</tr>
<tr>
<td>84</td>
<td>Correlation</td>
<td>93</td>
<td>Correlation</td>
</tr>
<tr>
<td>85</td>
<td>Correlation</td>
<td>94</td>
<td>Correlation</td>
</tr>
<tr>
<td>86</td>
<td>Correlation</td>
<td>95</td>
<td>Correlation</td>
</tr>
<tr>
<td>87</td>
<td>Correlation</td>
<td>96</td>
<td>Correlation</td>
</tr>
<tr>
<td>88</td>
<td>Correlation</td>
<td>97</td>
<td>Correlation</td>
</tr>
<tr>
<td>89</td>
<td>Correlation</td>
<td>98</td>
<td>Correlation</td>
</tr>
<tr>
<td>90</td>
<td>Correlation</td>
<td>99</td>
<td>Correlation</td>
</tr>
<tr>
<td>91</td>
<td>Correlation</td>
<td>100</td>
<td>Correlation</td>
</tr>
</tbody>
</table>
**Key word Analysis**

Cluster Analysis

Cluster analysis allows the user to generate a list of the most frequent 2-, 3-, 4-, 5-, or 6-word combinations (n-grams, word/lexical clusters/bundles) from a corpus, i.e. collocates.

(O’Keefe et al., 2001)

- Example application: Natural language processing (Part-of-Speech tagging), lexicography, study of formulaic language

**Cluster Analysis**

<table>
<thead>
<tr>
<th>Concordance</th>
<th>Concordance Flat</th>
<th>File View</th>
<th>N-grams</th>
<th>Collocates</th>
<th>Word List</th>
<th>Keyword List</th>
</tr>
</thead>
</table>

- From Chambers-Rostand
  - corpus: Oxford Text
  - Archive using Ant Conc.

**Concgramming**

- A ‘concram’ is all of the permutations of constituency variation and positional variation generated by the association of two or more words” (Greaves and Warren, 2007: 290)
- The words may be separated by a number of words
- The words may appear in any order
- Permit the identification of meaningful word associations within a corpus, that is the ‘aboutness’ of a corpus or its ‘phraseological profile’ (Greaves and Warren, 2007)

**Lexico-Grammatical Profiling**

- Collocates
  - Which word(s) occur most frequently and with statistical significance in the word’s environment?

- Chunks idioms
  - Does the word form part of any recurrent chunks? Is the word idiom-prone?

- Syntactic restrictions
  - Are there syntactic patterns which restrict the word? For example, are there prepositions that go with the word? What are its typical clause-positions (initial/mediatalfinal)? Are there any tense/aspect restrictions?

(Greaves and Warren, 2007: 299)

(O’Keefe et al., 2001: 14-15)
Lexico-Grammatical Profiling

- Semantic restrictions
  - Are there any semantic restrictions? For example, the word/phrases is applied to humans only, or is never used with an intensifier.
- Semantic prosody (Louw, 1993)
  - What are the connotative and attitudinal meanings of the word? Is the word positive or negative?
  - The collocates of cause are negative (accident, cancer, commotion)
  - The collocates of provide are positive (care, feed, help, jobs)

(O’Keeffe et al., 2001: 14-15)

Corpora in language learning

- Reference corpora
- Learner corpora
- Data-driven language learning

Reference Corpora

- Applications
  - Word lists
e.g. Academic Word List (Coxhead, 1998)
  - (Learner) dictionaries
e.g. Collins COBUILD English Language Dictionary
  - Grammars
e.g. Cambridge Grammar of English (Carter and McCarthy, 2006)
  - Textbooks and syllabies
e.g. The Touchstone series

Reference Corpora

For
- “...many features of real, naturally-occurring, spoken standard English grammar ... are not recorded in standard grammars of the English language” (Carter, 1998) e.g. three-part exchanges, vague language, ellipsis, formulaic language
- “The major standard grammars are ... Based largely on the written language and on examples drawn from single-sentence, sometimes concocted, written examples” (Carter, 1998)

Against
- “...computer corpora are incomplete. They contain information about production but not about reception. They say nothing about how many people have read or heard a text or utterance, or how many times ... Some phrases pass unnoticed precisely because of their frequency, others strike and stay in the min, though they may occur only once.” (Cook, 1998: 58)

Reference Corpora (cont.)

- “Corpora are records of language behaviour. The patterns which emerge in that behaviour do not necessarily and directly tell us how people organize and classify language in their own minds and for their own use, or how language is best systematized for teaching” (Cook, 1998: 58)
- “Even a three hundred million word corpus is equivalent to only around three thousand books, or perhaps the language experience of a teenager” (Cook, 1998: 59)
- “Native speakers acquire, represent, and process language in lexicalized chunks as well as grammar rules and single words. Yet it by no means follows that foreign learners must do the same” (Cook, 1998: 60)

Compromise
- “One conclusion reached so far in the preparation of discourse grammar materials is that a middle ground between authentic and concocted data might be occupied which involves modelling data on authentic patterns.” (Carter, 1998: 52)

Learner Corpora

- Applications
  - FreeText: A Smart Multimedia Web-based Computer-Assisted Language Learning Environment for Learners of French
    “FreeText offers four tutorials containing 16 authentic documents, ranging from texts to audiovisual files, which illustrate different communication acts. The exercises exploit these documents are based on studies of a learner corpus called FRIDA ... in order to concentrate on errors actually made by the target audience” (C’Haere and Vandevenne Falbin, 2003:482)
Learner Corpora

- **Corpora**
  - ICLE (Granger et al., 2002)
    - International Corpus of Learner English
    - Error tagged corpus of 2 million words of writing by learners of English from 19 different L1 backgrounds
  - FRIEDA (Granger et al., 2001)
    - Error tagged corpus of 450,000 words from essays written by French learners
  - Talkback project [www.talkback.org](http://www.talkback.org) (MacWhinney 2007)
    - 12 French, Spanish, Dutch, English, Welsh, Hebrew
    - Tagged spoken corpora with attached sound files.
    - Also contains speech data from patients with dementia and aphasia, as well as corpora coded for gesture etc.

Example activity with learner corpora: Developing more complex speech

- **FLLOC** ([www.flloc.soton.ac.uk](http://www.flloc.soton.ac.uk))
- Semi-elicited data with learners and native speakers
- Loch Ness story (LingDev, Newcastle corpora)
- Give class a selection of transcripts from different year groups (e.g. year 9-13, native speakers)
- Ask class to divide the transcripts according to proficiency.
- What clues did they use to categorize them?
- Lead into traditional exercises in use of discourse markers, connectives.
- Write their own story.

Data-Driven Language Learning

DDL, as described by Tim Johns, is intended to "confront the learner as directly as possible with the data, and to make the learner a linguistic researcher [...] [someone who is able] to recognize and draw conclusions from clues in the data [...]" (Johns, 2002: 108).

Vocabulary learning

- Compleat lexical tutor ([www.lexicon.ca](http://www.lexicon.ca))
  - ListLearn – vocabulary lists for French and English divided into frequency bands of 1000 words. Links to audio, concordance and dictionary.
  - Hypertext – upload your own text. Links with concordance, audio and dictionary. Reading resources for words.
  - Concordances – English, French, German, Spanish (soon)
  - Clone builder – upload your own document then decide what words to delete (e.g. every 5th word). Links to concordance.
  - N-gram – upload text and search for 3,4,5 word strings (useful for formulaic language)
  - Works best with Internet Explorer
Vocabulary/grammar

- Using AntConc
- [http://www.antlab.sci.waseda.ac.jp/software.html](http://www.antlab.sci.waseda.ac.jp/software.html)
- Concordance lines (edited)
- Option 1: give list of sentences with unknown word – what does it mean?
- Option 2: replace keyword with blank
- Option 3: distribution of two L2 words with same L1 meaning, e.g. to know (savoir vs. connaître)
- Option 4: Idiomatic uses of word

Reading/Listening

- Using Youth corpus (French, Spanish, German, Italian, Lithuanian, Romanian, English)
- [http://www.um.es/sacodeyl/](http://www.um.es/sacodeyl/)
- Based on videoed speech data
- Transcripts and resources available
- Searchable by topic, grammatical function etc.
- Students aged 11-18

Web as Corpus

- In the strictest sense of the term the Web is not a corpus – it is not balanced in any way

- Advantages:
  - "constantly expanding, self-renewing machine-readable body of linguistic data, much richer in current language usage, infrequent expressions, text genres and domains than even the biggest standard reference corpus" (Krajka, 2009: 418)...
  - "freshness and spontaneity, completeness and scope, linguistic diversity, representativeness and free availability" (ibid.)

- Disadvantages:
  - "huge rag bag of digital text" (Krajka, 2009: 418)
  - Unmodified, non-native, etc.
  - Teachers need to carefully select their corpus (Robb, 2003)

Savoir versus connaître

à lui. Mais moi, malheureusement, je ne sais pas voir les mouettes à travers les caill... sans doute, il sait tout le monde au soleil, il couche sur la France (suffisamment) de nuit. Mais, comme il disait, on ne sait jamais ! Il ramena donc également le seul... mais il n’a personne à jouer ? On ne sait pas, lui dit-il. On n’a pas fait à... j’aimerais de travailler ! On ne sait plus... le sais-tu, moi, je ne l’aime pas il y a des années. Mais en ne sait jamais où les trouver. Le vent les prom... avec de moi. Quelle promenade ! Te vois... une muscadelle pour mon mouton... il envoie un effort. Il a été génial, tu vois. Moi aussi je regarde les oiseaux. Il y... et à être... pas tout à fait. Mais, je suis bien qu’il est revenu à sa pensée, car... vole semblable à quelque part, en sait où éléphant que nous ne connaissions pas au vent des chevaux tout d’un seul : le connaissons une plante où il y a un monstre comme... d’un gros monstre noir. Il est le connaissez, moi, une fleur unique au monde, qu’t’as... tu sais !. Il veut aider son ami. Je connais un moyen de te repérer quand tu souffres un peu, sa dit que, en un peu ou une seule. Le connaissez quelqu’un, dit le petit prince, qui ser... idée de notre plante à qui le ne le connaissait pas. Les hommes occupent tous peu de pl... un ne sait où, un mouton que nous ne connaissons pas, a, lui en même, mangé une rose... 

Extracts from Le Petit Prince [http://www.unfoundation.org/lsp/sentences.txt](http://www.unfoundation.org/lsp/sentences.txt)

‘From textbook to data’ or ‘from data to textbook’?

"The principle of fidelity to the data is one which we ignore at our, and our students’, peril. That danger is well illustrated by Groß, Müller and Wolff (1996), which uses concordance data to teach the old textbook rule for the use of some and any in English: some in positive statements, any in negative statements and in questions. Reference to any (I) KWIC concordance of any will show that generalisation to be false: the problem is that having decided on the generalisation in advance, it is all too easy to select only those citations that support it”

(Johns, 2002).
Web as Corpus

Research

- Boulton (2007)
  - Reviewed 19 empirical papers on DOL
  - In 34 studies English was the target language
  - Only 2 studies focused on younger learners
  - 36 studies were conducted in higher education institutes
    - 33 focused on language learning, 3 focused on linguistics
    - Only 2 claim “low” levels and 2 “beginners”
    - A variety of corpora were used (Bank of English, British National Corpus, ICE, MCAST, custom)
    - In most studies allowed directly accessed corpora using WordSmithTools
    - RQs: (1) Attitudes, (2) learners’ practices, (3) learning outcomes
    - Only 6 evaluate learning outcomes – these focus on lexicological
  - Results: “Learners attitudes are largely positive; in most cases they are remarkably capable of corpus techniques; corpora can be used as an effective reference tool, as well as for learning” (Boulton, 2007: 14)

Research

- Chambers (2007)
  - Quantitative
    - Steenev (1991): Concordance-based exercises on paper better than gap-filler exercises for vocabulary acquisition
    - Cobb (1997): On-screen concordance-based exercises are better than the use of other resources
  - Qualitative
    - Positive
      - “Appreciate the relevance of the corpus data” (Bernadini, 2002)
      - “Provide examples of language ‘in context’” (Yoon and Hirvela, 2004; Chambers and O’Sullivan, 2004)
      - “Appreciate the abundance of examples” in comparison with a dictionary (Cheng et al., 2003; Yoon and Hirvela, 2004; Chambers, 2005)
      - Appreciate the self-directed nature of DOL (Bernadini, 2002; Chambers, 2005)
      - Find the activity motivating (Chambers, 2005)
Research

- Chambers (2007)
  - Qualitative
  - Negative
  - Difficult (Cheng et al., 2003)
  - Time-consuming (Yuan and Hirvela, 2004; Chambers and O’Sullivan, 2004)
  - Laborious and tedious (Cheng et al., 2003; Chambers, 2005)
  - Frustrating (Bernadin, 2000; Cheng et al., 2003; Chambers and O’Sullivan, 2004)
  - Learners require training (Bernadin, 2002; Cheng et al., 2003; Chambers and O’Sullivan, 2004; Gaskell and Cobb, 2004; Chambers, 2005)

Research

*At this early stage in the development of corpus consultation by learners, qualitative information, alongside quantitative studies, is undoubtedly useful for other researchers and practitioners involved in similar activities, who can learn from accounts of what others have done, of what has worked well and what problems have been encountered*  
(Chambers, 2007: 7)

*Given the number of variables involved, no single study is likely to ‘prove’ very much, just as a single concordance line is not the best evidence for language use. To take the analogy further, corpus linguistics looks at many concordances to find the general tendencies of language patterning; what is needed here is a large number of studies in DOL to see where the weight of evidence takes us. Without empirical support, the most we can hope for are statements along the lines of ‘I think’, ‘It seems to me’, ‘In our opinion’, etc. – which do indeed feature prominently in the DOL literature*  
(Boulton, 2007: 14)

Theory

- Vocabulary knowledge
  - Form: spoken, written, word parts
  - Meaning: form and meaning, concept and referents, associations
  - Use: grammatical functions, collocations, constraints on use (register, freq)

  (Nation, 2001)

- Ideal psychological conditions for vocabulary learning
  - Noticing
  - Comprehension
  - Retrieval
  - Generative use

  (Nation, 2001)

Benefits

- Automatic searching and sorting (Leech, 1997)
- Open-ended supply of language data (Leech, 1997)
- Enables the learning process to be tailored (Leech, 1997)
- Authentic language
- Promotes a learner-centred approach (Leech, 1997)
- Learner autonomy (Chambers and Kelly, 2002)
- Processing authentic texts can increase learners’ metalinguistic knowledge (Gavioli, 1997)
- Engaging and “something different”

Limitations

- Volume of information may overwhelm students (Cobb, 1998) or teachers
- Unknown words in the contexts (Cobb, 1998)
- Contexts are short and incomplete (Cobb, 1998)
- Required training for efficient use (Stevens, 1995)
- Learners may treat the corpus as another dictionary (Stevens, 1995)
- Not all learners have positive attitudes to inductive learning (Krieger, 2003)
- Difficulty of assessing such an open-ended task (Leech, 1997)

Working within the Limitations

- Simplify the data
  - Select familiar/predictable data
  - Reduce the quantity of data
- Simplify the task
  - Recognition vs. induction
  - Predetermined categories vs. devising categories
  - Group work vs. individual work
  (Aston, 1997)
- Use print-outs/interactive whiteboard
  (Johns)
Recommended Reading


[http://www.lexically.net/words/nth/corpus_4perience_links/Tim%20johns%20and%20DL.pdf] [Data-driven language learning in the words of Tim Johns]


[http://ilt.msu.edu/col8num3/pdf/review2.pdf] [Review of free concordancing tools for language learning]

Other References


Materials


Resources and Tools

Vivienne Rogers & Zoe Handley

Webcorp: [http://www.webcorp.org.uk/]

Output: KWIC for web pages from selected domains / so multilingual

Options

• Search engine
• Case sensitivity
• Output format (plain/HTML)
• Web addresses for corpus lines
• Concordance span [no. words left and right] / whole sentences
• Number of pages to retrieve
• Site domain [e.g. .ac.uk]
• Textual domain (topic)
• Word filter [extra words which must (not) appear in the concordance lines]

Youth Corpus: [http://www.um.es/sacodeyl/]

• Available in English, French, Spanish, Italian, Lithuanian and Romanian
• Video and transcripts of interviews with 11-18 year olds.
• Fixed range of topics/categories including family, free time, elections etc.
• Search engine (by word, topic or grammatical feature)
• Teacher/learner resources available for some topics
• Requires RealPlayer for playback of video clips.

AntConc [http://www.antlab.sci.waseda.ac.jp/software.html]

Description: [http://www.antlab.sci.waseda.ac.jp/research/lwefl_2004_anthony_antconc.pdf]

Options/features:

• Use own corpus
• Concordancer
• Frequency lists
• Keyword generator
• Cluster and lexical bundle analysis

Compleat lexical tutor [http://www.lextutor.ca]

French, English Spanish

Options/features:

• Frequency > Word profile
• Range > Word profile + (Text_lex_compare) recycling index
• Concordancer (French, German, English a range of corpora + Custom for French and English)
• Story concordancer – links words to concordances from the story
• N-Gram Phrase Extractor (2 – 5 words)
• VocabProfile (Word profile)
• TextLexCompare (Compare word profile with word lists)
• Links reading text to concordancer and WordNet (English only)
• KeyWords Extractor (English only)
• Multi_Conc (Multiple choice concordancer – English only)
• ID Word Identification Quiz (Guess the word that fills the concordance – English only)
• Cloze Builder (English and French)
• HyperTest (Put in own text and link to concordances to help with guessing for further contexts)


Options/features:

• Use own corpus / custom corpus (untagged/annotated)

Vivienne Rogers & Zoe Handley
• Phrase, word, prefix, suffix searches
• Concordancer
• Collocations
• Conigramming
• WordProfile
• Compare word profile with word lists (2 lists provided English)
• Link to Net Dictionary

TextSTAT
Download: http://neon.niederlandistik.fu-berlin.de/textstat/
User guide: http://sites.google.com/site/gernabennett2/TextSTATusersguide.pdf?attredirects=0
Options/Features:
• User compiled corpus creation
• Frequency lists for the corpus
• Keyword searches
• Wild card searches
• Searches for two expressions with a number of words between them

Simple Concordance Program (http://www.textworld.com)
Options/features:
• Use own corpus
• Keyword search
• Prefix/suffix/anywhere search
• Case sensitivity
• Frequency lists
• Word profile

Corpora
French
• Chambers-Rostand Corpus of Journalistic French: http://www.otd.ahds.ac.uk/%20texts/2491.html
• Lexicometrie: Corpus of classic French literary texts
  http://ota.ahds.ac.uk/headers/3466.xml
• Le Corpus BAF (English-French Parallel): http://rali.iro.umontreal.ca/

Italian
• Banca dai dell'italiano parlato (BADIP): http://languageserver.uni-graz.at/badip/
• Corpus di Italiano Scritto (CORIS): http://corpus.cltta.unibo.it-8080/CORISCorpQuery.html

Spanish
• Corpus Oral de Referencia del Espanol Contempanoeo (COREC):
  http://www.lifl.uam.es/corpus/corpus_oral.html (Sample:
  http://www.lifl.uam.es/corpus/corpus_lee.html#4)
• The CREA Corpus of Spanish: http://www.rea.es/ AND http://corpus.cae.es/creanet.html

Multilingual
• TRACTOR archive: http://www.corpus.bham.ac.uk/ccl/services.htm#tractor

Corpus-based dictionaries
French

German
• Elsiko: www.elsiko.de
• Neuman et al. (DATE). A Corpus-based lexical resource of German idioms. Produced as part of the “Collocations in the German Language” project at Berlin-Brandenburg Academy of Sciences.

Multilingual
• Worschatz Universitat Leipzig: http://corpora.informatik.uni-leipzig.de/?dict=es
• >> Search 59 Corpus-based monolingual dictionaries

Corpus-based Curriculare