# Describing target language domains with #LancsBox

ⓘ In this session, you will be using **#LancsBox** (http://corpora.lancs.ac.uk/lancsbox) and **Lancaster Stats Tools online** (http://corpora.lancs.ac.uk/stats)

More information:
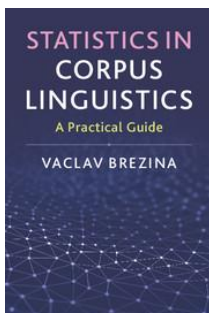
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139-173.
- Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

---

Ⓣ **Task 1.** Start #LancsBox and load the LOB corpus.

#LancsBox is a free software tool developed at Lancaster University (lead developer: Dr. Vaclav Brezina). It can be downloaded and used for free on any major operating system (Windows, Mac and Linux). #LancsBox has already been pre-installed on the computers in the computer lab.

1. Go to 'AppsAnywhere' **lancaster.ac.uk/apps**
2. Scroll-down to LancsBox (under L).
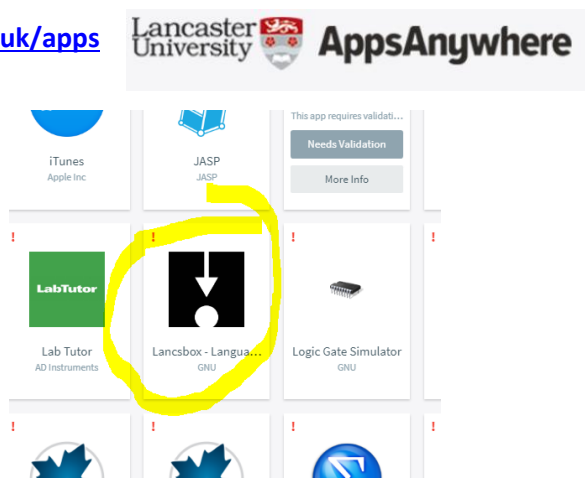
3. Run #LancsBox by clicking on the icon and launch the tool.

4. Load the LOB corpus, a one-million-word sample of British writing, by following the steps below.

❶ On the Corpora tab under 'Download' click on 'Corpus'.

❷ From the list select 'L-O-B'.

❸ Click on 'Download'

❹ and then on 'Import'.

❺ Wait while the corpus is being processed (notice the progress bar at the bottom of #LancsBox window).
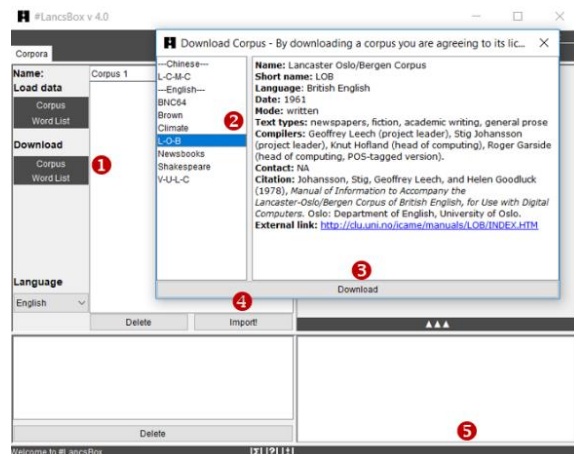
**Task 2.** Use the 'Whelk' tool to search the LOB corpus. Note down the frequencies of words and phrases as well as their distributions in different genres/registers of written English.
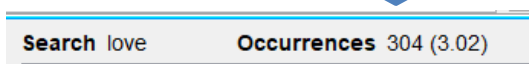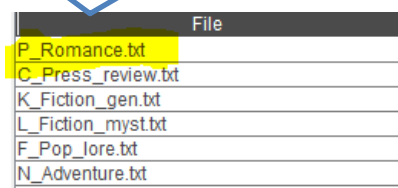
1. From the main #LancsBox menu select 'Whelk'.
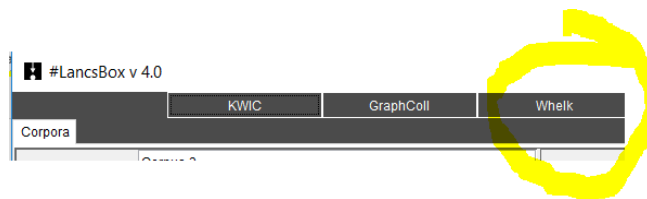
2. Search the LOB corpus. Type in the search terms from the table below into the search box.

| Search term | Occurrences (per 10k) | Genres/registers in which the search term is frequent |
|---|---|---|
| Love | | |
| Research | | |
| research* | | |
| research and * | | |
| VERBS | | |
| PASSIVES | | |

Look at the number next to the word 'Occurrences'. The number in parentheses () represents the frequency per 10k running words.

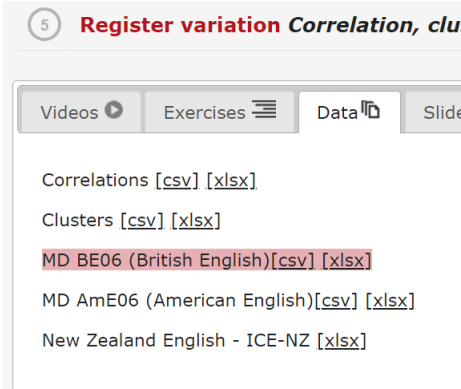Look at the list of genres – the search term is most frequent in the ones on the top.

**Search** love    **Occurrences** 304 (3.02)

| File |
|---|
| P_Romance.txt |
| C_Press_review.txt |
| K_Fiction_gen.txt |
| L_Fiction_myst.txt |
| F_Pop_lore.txt |
| N_Adventure.txt |

(T) **Task 3.** Think about which linguistic features could be used to distinguish different genres/registers. Look at a list of variables below that have been shown (Biber 1988) to occur with different frequencies in different genres/registers. Which ones are most important for distinguishing genres/registers?

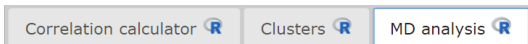| | | |
|---|---|---|
| 1. past tense | 24. infinitives | 47. hedges (e.g., *almost, maybe, sort of* [except as true noun]) |
| 2. perfect aspect | 25. present participial adverbial clauses (e. g., *Screaming with rage, he ran up the stairs.*) | 48. amplifiers (e.g., *completely, totally, utterly*) |
| 3. present tense | 26. past participial adverbial clauses (e.g., *Given these characteristics, it is not surprising that...*) | 49. emphatics (e.g., *a lot, for sure, really*) |
| 4. place adverbials (e.g., *behind, downstairs, locally*) | 27. past participial postnominal (reduced relative) clauses (e. g., *the exhaust air volume required by the 6-ft. x 4-ft. grid*) | 50. discourse particles (e. g., sentence initial *anyhow, now, well*) |
| 5. time adverbials (e.g., *eventually, immediately, nowadays*) | 28: present participial postnominal (reduced relative) clauses (e.g., *the currents of dissent swirling beneath the surface*) | 51. demonstratives |
| 6. first-person pronouns | 29. *that* relative clauses in subject position (e. g., *the papers that are on the table*) | 52. possibility modals (*can, could, may, might*) |
| 7. second-person pronouns | 30. *that* relative clauses on object position (e.g., *the papers that she thought would be interesting*) | 53. necessity modals (*must, ought, should*) |
| 8. third-person personal pronouns (excluding *it*) | 31. *Wh*-relatives in subject. position (e. g., *people who know him*) | 54. predictive modals (*shall, will, would*) |
| 9. pronoun *it* | 32. *Wh*-relatives on object position (e. g., *people who he knows*) | 55. public verbs (e. g., *complain, explain, promise*) |
| 10. demonstrative pronouns (*that, this, these, those* as pronouns) | 33. pied-piping relative clauses (e. g., *the way in which food is digested*) | 56. private verbs (e.g., *believe, think, know*) |
| 11. indefinite pronouns (e.g., *anyone, everybody, nothing*) | 34. sentence relatives (e.g., *We waited for six hours, which was ridiculous.*) | 57. suasive verbs (e.g., *command, propose, recommend*) |
| 12. pro-verb *do* | 35. causative adverbial subordinator (*because*) | 58. *seem* and *appear* |
| 13. direct Wh-questions | 36. concessive adverbial subordinators (*although, though*) | 59. contractions (*don't*) |
| 14. nominalizations (ending in -*tion*, -*ment*, -*ness*, -*ity*) | 37. conditional adverbial subordinators (*if, unless*) | 60. complementizer *that* deletion (e.g., *I think [Ø] he's gone already.*) |
| 15. gerunds (participial forms functioning as nouns) | 38. other adverbial subordinators (e.g., *insomuch as, such that, while*) | 61. stranded prepositions (e.g., *the person that I was talking to*) |
| 16. total other nouns | 39. total prepositional phrases | 62. split infinitives (e.g., *I want to completely convince you that* |
| 17. agentless passives | 40. attributive adjectives (e.g., *the small room*) | 63. split auxiliaries (e.g., *they have apparently sold it all ...*) |
| 18. by-passives | 41. predicative adjectives (e.g., *the room is small*) | 64. phrasal coordination (NOUN and NOUN; ADJ and ADJ; VERB and VERB; ADV and ADV) |
| 19. *be* as main verb | 42. total adverbs | 65. independent clause coordination (clause initial *and*) |
| 20. existential *there* | 43. type/token ratio | 66. synthetic negation (e.g., *No evidence was found.*) |
| 21. *that* verb complements (e.g., *We felt that we needed a financial base.*) | 44. mean word length | 67. analytic negation (e.g., *That's not true.*) |
| 22. *that* adjective complements (e.g., *It's quite obvious that certain things can be sexlinked.*) | 45. conjuncts (e. g., *alternatively, nevertheless, therefore*) | |
| 23. Wh-clauses (e. g., *I wondered what to do.*) | 46. downtoners (e.g., *mildly, partially, somewhat*) | |

**T** **Task 4.** Perform the multidimensional analysis of registers in BE06, a one-million-word corpus of current British English, using the Multidimensional analysis tool from Lancaster Stats tools online.

      a) Which linguistic features appear together in the same dimension?

      b) Which genres/registers cluster together?

Follow the steps below.

1. Download the MD BE06 dataset from
   http://corpora.lancs.ac.uk/stats/materials.php?panel=4&tab=2 and open it in Excel.



2. Select (Ctrl+A), copy (Ctrl+C) and paste (Ctrl + V) into the MD tool
   (http://corpora.lancs.ac.uk/stats/toolbox.php?panel=4&tab=2 ).

3. Choose 'Full MD'.



4. Press the 'Perform MD analysis' button.

5. Select 5 factors to extract.
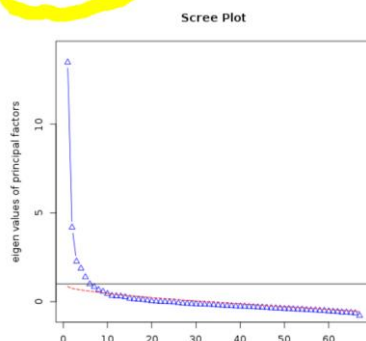
6. Interpret the results.

FACTOR ANALYSIS

| Features | Factor1 ▲ | Factor2 | Factor3 | Factor4 | Factor5 |
|---|---|---|---|---|---|
| CONT | 0.934 | -0.171 | -0.329 | 0.050 | 0.004 |
| ANNEG | 0.861 | 0.083 | 0.006 | -0.052 | 0.123 |
| PRIV | 0.729 | 0.252 | -0.001 | -0.182 | 0.076 |
| X2PRON | 0.717 | -0.064 | -0.288 | 0.216 | -0.049 |
| STPR | 0.684 | -0.123 | -0.134 | -0.099 | 0.055 |
| BE | 0.648 | -0.004 | 0.165 | 0.127 | 0.154 |

Sort the table by clicking on the column labels e.g. 'Factor 1', 'Factor 2' etc.. to see which variables are strongly present (numbers close to 1) or strongly absent (numbers close to -1) in each factor dimension. Refer back to the features in Task 3.
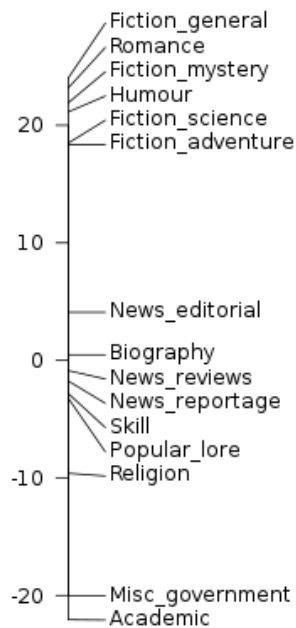


Look at the dimension scale to see which genres/registers are similar (close together) and which are different (further apart).

# Notes