2. (2339 + 2089 + 2056 + 2276 + 2233 + 2056 + 2241 + 1995 + 2043 + 1976 + 2062) / 11 = 2124.18

3. model in scientific thinking means: model is a way to look at a data set and will help us notice some features of the data set that we cannot see if we look at the data individually. It is a simplified representation of the data but preserves all the features of the original data set. In corpus linguistics, model is mathematically representing linguistic reality.

4. (C) triangle

5. (520 \* 900) /2 = 234000 km2

6. (i) B (ii) A (iii) B (iv) B (v)A (vi) C (vii) C (viii) B

7. 161, 415, 390, 455, 294, 280, 400, 265, 13, 2, 430, 447, 126, 398, 256, 407, 160, 271, 404, 91, 453, 287, 93, 465, 474, 225, 212, 245, 445, 466, 17, 209, 36, 193, 186, 362, 374, 328, 11, 51

8. male: 23, female: 17 This is not exactly equal but approximately equal. The equal expected should be 20 and 20.

9. There are 21 old speakers and 19 young speakers. To be specific, there are 11 male old speakers, 10 female old speakers, 12 male young speakers, and 7 female young speakers.

10. This is not a good method. The selected samples do not represent the actual ratio of the original population. There is equal amount of male and female, old and young speakers, but this selected sample doesn’t have equal numbers in each category, thus it fails to capture this feature of the original dataset.

11. We should avoid text sample bias, topic bias, non-coverage bias, traditional text type bias, legal consideration bias, practicality bias, self-selection bias,

12. (i) Linguistic Feature (ii) Individual text, speaker design (iii) whole corpus (iv) individual text, speaker design

13. the line of the, the line of and the, the frequency per million for to, and total

14. (a) histogram (b) box plot (c) scatter plot (d) error bars

15.