Chapter 2: Exercises – answers

1) Identify the no. of *tokens, types, lemmas* and *lexemes*.

a)

Tokens (26)	Types (23[1])	Lemmas (23[2])	Lexemes (23)
The; City; is; braced; for; far; worse; figures; to; come; in; the; coming; months; unless; the; Government; recovery; package; produces; a; startling; turn; round; in; optimism	the; city; is; braced; for; far; worse; figures; to; come; in; coming; months; unless; government; recovery; package; produces; a; startling; turn; round; optimism	the; City; be; brace; for; far; bad; figure; to; come; in; coming; month; unless; Government; recovery; package; produce; a; startling; turn; round; optimism	THE; CITY; BE; BRACE; FOR; FAR; BAD; FIGURE; TO; COME; IN; COMING; MONTH; UNLESS; GOVERNMENT; RECOVERY; PACKAGE; PRODUCE; A; STARTLING; TURN; ROUND; OPTIMISM

b)

Tokens (29[3])	Types (27[4])	Lemmas (24[5])	Lexemes (24)
Of; 354; fifth-; and; sixth-formers; who; left; Sharon's; school; in; the; summer; of; 1981; forty; had; found; real; jobs; by; 18; November; four; of; these; having; entered; military; service	of; 354; fifth-; and; sixth-formers; who; left; sharon's; school; in; the; summer; 1981; forty; had; found; real; jobs; by; 18; november; four; these; having; entered; military; service	Of; <NUMBER>; fifth-; and; sixth-formers; who; leave; Sharon; school; in; the; summer; forty; have; find; real; job; by; November; four; these; enter; military; service	OF; <NUMBER>; FIFTH-; AND; SIXTH-FORMERS; WHO; LEAVE; SHARON; SCHOOL; IN; THE; SUMMER; FORTY; HAVE; FIND; REAL; JOB; BY; NOVEMBER; FOUR; THESE; ENTER; MILITARY; SERVICE

[1] An alternative solution: 24 if the case sensitive option is selected – *The* and *the* would be counted as two types.

[2] Alternative solutions: a) 22 if *turn round* is understood as one lexical unit b) 22 if *coming* is lumped under the headword *come*.

[3] An alternative solution: 30 if hyphen considered as a token separator; in that case *sixth* and *formers* would be considered as two tokens.

[4] An alternative solution: 28 if the case sensitive option is selected – *Of* and *of* would be counted as two types.

[5] An alternative solution: 25 if possessive suffix *'s* is counted as a separate lemma.

c)

Tokens (14)	Types (12[6])	Lemmas (12)	Lexemes (10[7])
Erm; erm; erm; but; yeah; and; people; er; have; great; areas; of; that; taken	erm; but; yeah; and; people; er; have; great; areas; of; that; taken	erm; but; yeah; and; people; er; have; great; area; of; that; take	BUT; YEAH; AND; PEOPLE; HAVE; GREAT; AREA; OF; THAT; TAKE

d) This is a very specific example which includes meta-linguistic comments on the meanings/uses of the form *bow.*

Tokens (26)	Types (18)	Lemmas (19)	Lexemes (20)
Homonyms; are; headwords; to; different; entries; that; are; spelt; in; the; same; way; e.g.; bow; the; weapon; bow; the; action; bow; the; verb; expressing; the; action	homonyms; are; headwords; to; different; entries; that; spelt; in; the; same; way; e.g.; bow; weapon; action; verb; expressing;	Homonyms; be; headword; to; different; entry; that; spell; in; the; same; way; e.g.; bow; weapon; action; bow; verb; expressing;	Homonyms; be; headword; to; different; entry; that; spell; in; the; same; way; e.g.; bow; weapon; bow; action; bow; verb; expressing;

2) and 3) –

4) Calculate the relative frequencies.

a) *muggle*: 0.2 per 10k

b) *intriguingly:* 0.3 per million

b) *worse:* 49.6 per million

[6] An alternative solution: 12 if the case sensitive option is selected – *Erm* and *erm* would be counted as two types.

[7] The paralinguistic hesitation sounds (erm and er) in this utterance from a transcript of spoken conversation were excluded because they do not have a semantic meaning.

5) Use *Zipf's law* to predict absolute frequencies.

rank	word	absolute frequency
1.	the	6,041,234
2.	of	3,020,617
3.	and	2,013,745
4.	to	1,510,309
5.	a	1,208,247
10.	was	604,123
50.	so	120,825
100.	way	60,412
1,000.	limited	6,041
10,000.	conveniently	604

6) N.B. Zipf's law is only an approximation and the actual absolute frequencies in the table below differ to some extent from the predicted ones.

rank	word	absolute frequency
1.	the	6,041,234
2.	of	3,042,376
3.	and	2,616,708
4.	to	2,593,729
5.	a	2,164,238
10.	was	881,473
50.	so	239,116
100.	way	95,701
1,000.	limited	10,312
10,000.	conveniently	622

7) Calculate the Range, the Standard deviation, the Coefficient of variation and Juilland's D.

Note that the first step is to convert all absolute frequencies to relative frequencies as seen in the table below.

BNC section	Total no. of tokens	*some* (RF)	*smile* (RF)	*theory* (RF)	*chance* (RF)
Fiction and verse	16,143,913	1,525	341	21	164
News-papers	9,412,174	1,118	32	28	275
Non-academic prose and biography	24,178,674	1,785	16	164	91
Academic prose	15,778,028	1,920	4	418	58
Other written material	22,390,782	1,691	22	57	148
Spoken	10,409,858	1,978	11	35	109

a) Range

some: 6

smile: 6

theory: 6

chance: 6

b) Standard deviation

some: 287.74

smile: 121.06

theory: 141.54

chance: 69.46

c) the Coefficient of variation

some: 0.17

smile: 1.71

theory: 1.17

chance: 0.49

d) Juilland's D

some: 0.92

smile: 0.24

theory: 0.47

chance: 0.78

8) Use *Juilland's U* usage coefficient to rank the words *some*, *smile*, *theory* and *chance* according to their relative importance.

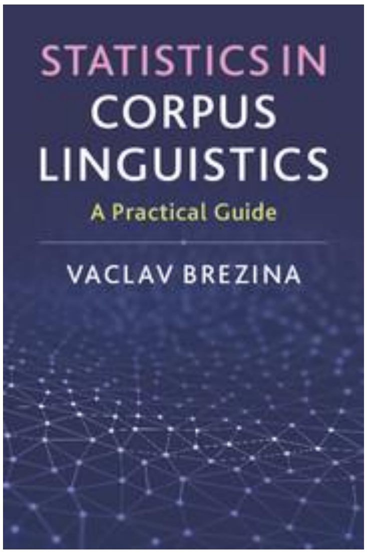
	Juilland's D	AF (whole corpus)	Juilland's U (Juilland's D × AF)
1. some	0.92	167,050	153,686.00
2. chance	0.78	12,809	9,991.02
3. theory	0.47	12,809	6,020.23
4. smile	0.24	6,848	1,643.52

9) Calculate the ARF of the selected words in the *BE06* corpus (985,628 tokens):

a) *frigid*: ARF = 1.02

b) *chemistry*: ARF = 3.17

c) *porn*: ARF = 4.6



Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press.

Do you use language corpora in your research or study, but find that you struggle with statistics? This practical introduction will equip you to understand the key principles of statistical thinking and apply these concepts to your own research, without the need for prior statistical knowledge. The book gives step-by-step guidance through the process of statistical analysis and provides multiple examples of how statistical techniques can be used to analyse and visualise linguistic data. It also includes a useful selection of discussion questions and exercises which you can use to check your understanding.

The book comes with a Companion website, which provides additional materials (answers to exercises, datasets, advanced materials, teaching slides etc.) and Lancaster Stats Tools online, a free click-and-analyse statistical tool for easy calculation of the statistical measures discussed in the book.