## Lecture 2. Vocabulary: Frequency, dispersion and diversity

Aim: The lecture is based on Brezina (2018), Chapter 2. It introduces simple statistical measures that help describe the occurrence of words in texts and corpora. It focuses on word frequencies and distributions both of which are crucial for meaningful description of patterns of language use.

- Key terms: token type lemma lexeme Zipf's law absolute frequency relative frequency dispersion Average reduced frequency type/token ratio

Time:

1-hour lecture.

2-hour computer lab session with exercises and Lancaster Stats Tools online (optional).

1-hour individual study (readings).

Statistical tools: Word calculator, Wordlist, Dispersion calculator and ARF Calculator

Practical exercises: Chapter 2 Exercises and answers.

Data: Chapter 2 data.

Readings: Brezina 2018 Chapter 2 + optionally selected Advanced readings recommended in the book.

Outline:

1. Tokens, types, lemmas and lexemes
2. Words in a frequency list
3. The whelk problem: Dispersion
4. Which words are important? Average reduced frequency
5. Lexical diversity: Type/token ratio (TTR), STTR and MATTR
6. Things to remember

Main points – 'Things to remember':

- There are different concepts of a 'word' – token, type, lemma and lexeme.
- Zipf's law describes the distribution of words in corpora and their rapidly diminishing frequency.
- To fully describe a word in a corpus we need to provide both the word's frequency and its dispersion.
- Different dispersion measures (Range, SD, CV, CV%, Juilland's D, DP) are appropriate in different situations.
- The average reduced frequency (ARF) is a measure that combines both frequency and dispersion; it can be used with corpora that are not divided into different parts (subcorpora).
- TTR is a measure of lexical diversity; it is sensitive to text length.
- STTR and MATTR are alternative measures of lexical diversity that can be used with texts of varying lengths.