Meeting 6:

Keywords and Probabilities

25 November 2021

Types and tokens

Token: total number of words in a corpus

Type: number of unique words in a corpus, double or multiple occurrence not counted

- Type-Token-Ratio(TTR): the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language;

  → the closer the TTR ratio is to 1, the greater the lexical richness of the segment

(Standardised Type-Token-Ratio (STTR): average of ratio for types divided through tokens for each 1000 words)

- Normalised Frequency (or relative frequency): comparison tool, upscale corpus tokens to 1,000 or 1,000,000, then compare counts of keyword per thousand / a million words

Definition

Keyword:

“A keyword refers to a lexical item which occurs with unusual frequency, either with a significantly higher or lower frequency in a target text or corpus, when compared to a reference corpus“

(Pojanapunya and Watson Todd, 2016, 1)

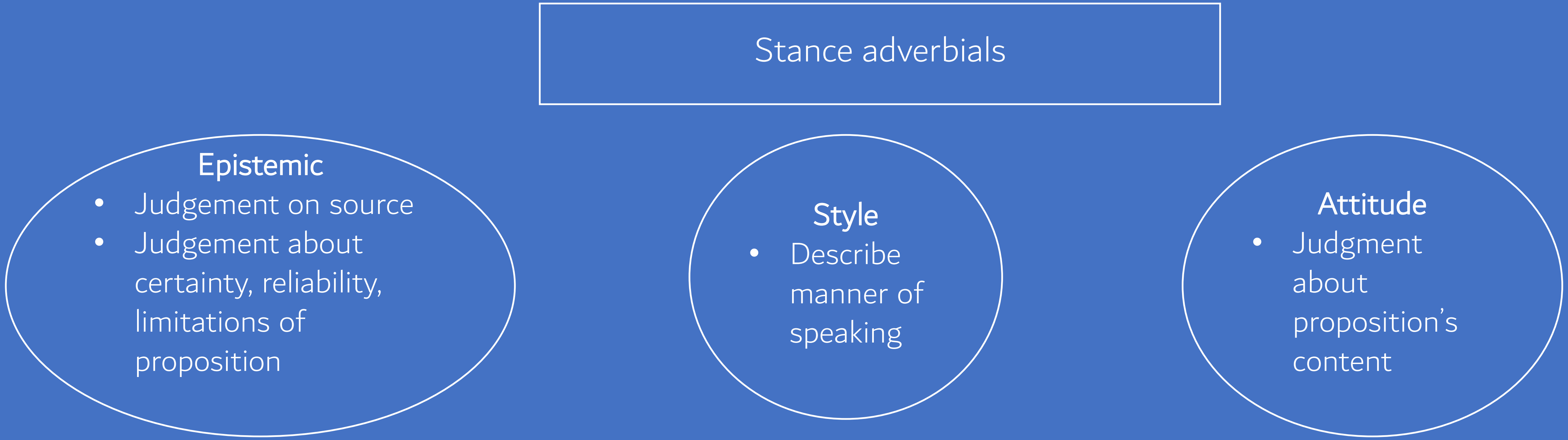
Definition

Homonym: a word that is spelled and pronounced like another word but is different in meaning

→ often: verb and noun

Definition

Stance Adverbs: Stance adverbs express the speaker or writer's point of view or judgment in relation to the particular circumstances associated with the content of a clause. They are language-, culture- and context-dependent (cf. Biber et al. 1999: 854).



Stance adverbials

Epistemic

- Judgement on source
- Judgement about certainty, reliability, limitations of proposition

Style

- Describe manner of speaking

Attitude

- Judgment about proposition's content

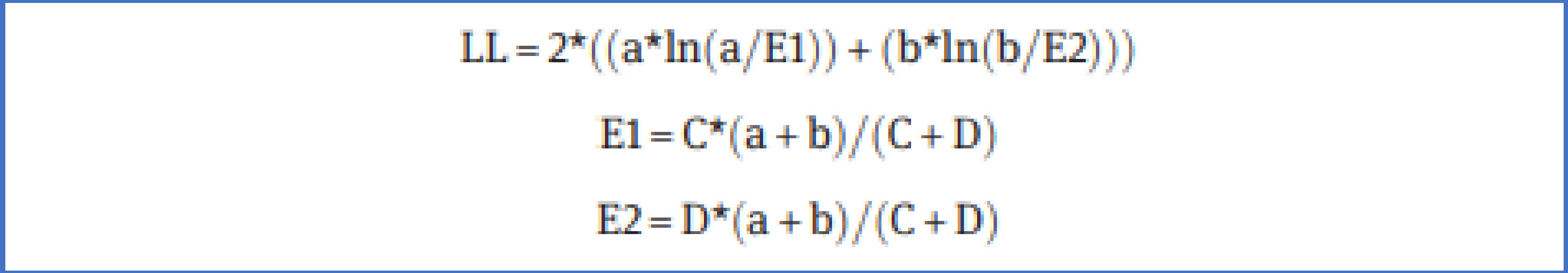
Stop words

https://www.ranks.nl/stopwords: Stop words for English

Attention: needs careful adaptation to your research needs

Log Likelihood

- high where there is a great disparity in frequency
- a simple version of the LL calculation is presented in Rayson (2013)



$$LL = 2*((a*\ln(a/E1)) + (b*\ln(b/E2)))$$

$$E1 = C*(a + b)/(C + D)$$

$$E2 = D*(a + b)/(C + D)$$

Online calculator: http://ucrel.lancs.ac.uk/llwizard.html

Group Work

Find out if “time” is a keyword in our example corpus (cf. email from last week) with help from AntConc and the Log Likelihood method.

Helpful links and numbers:

http://ucrel.lancs.ac.uk/llwizard.html

https://www.laurenceanthony.net/software/antconc/

https://www.english-corpora.org/bnc/

General corpus BNC: 100,000,000 words

Further Reading

Please have a look at this website and article for further statistic understanding:

McEnery, Tony/ Hardie, Andrew (2021): Statistics in corpus linguistics. Corpus Linguistics: Method, theory and practice. Lancaster University. Available at http://corpora.lancs.ac.uk/clmtp/2-stat.php.

Pojanapunya, Punjaporn/Watson Todd, Richard (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Linguistic Theory 14 (1), 133-167. https://doi.org/10.1515/cllt-2015-0030.