Visual evaluation framework for specialised corpora

Corpus linguists use deductive reasoning to attempt to falsify null hypotheses to predetermined levels of statistical significance. Conclusions are inferred from the corpus results using inductive reasoning, such as generalization, simple induction, argument from analogy and causal inference. These conclusions, however, may be rejected for various reasons, such as sample bias. Avoiding bias in corpora is therefore a primary aim of corpus compilers.

Corpora that are unbiased may be described as representative. Representativeness is a nebulous notion that is inextricably intertwined with sampling, size and balance. Providing sufficient supporting evidence is more convincing than simply claiming representativeness. However, in many research articles, there is often insufficient data to attempt to verify representativeness. Scant descriptions of the population, incomplete sampling frames and a dearth of data on central tendencies and standard deviation within corpora are cases in point. However, statistical descriptions alone may not resolve the problem, since applied linguists tend to be ill at ease with statistical concepts and terminology.

This paper identifies the often overlooked assumptions in the design of specialised corpora, highlights the logical fallacies to be aware of and helps linguists navigate through vague and ambiguous statistical jargon. A visual evaluation framework is proposed to provide a reader-friendly way for corpus compilers to show that the corpus design criteria are sufficient for the intended purpose.

Specifically, sampling decisions, such as defining the population, sampling frame and unit, are clarified. Venn diagrams depict the relative proportions of the corpus and its population. Empirical analysis of a preliminary corpus is used to graph cumulative variation and predict optimal size. Balance is illustrated using pie charts that show the strata within corpora and its population. Internal and external consistency in terms of central tendencies and standard deviation are visualized using box and whisker plots, created from summary statistics of subsampling and cross-validation data sets.

This visual evaluation framework presents information graphically, numerically, algebraically and verbally. This enables readers to more easily understand the representativeness of the corpus in relation to its purpose. A case study of the creation of a specialised corpus of scientific research abstracts is used as a vehicle to exemplify the usefulness of this framework.