Appendix 1: Fundamental Principles of Corpus Linguistics

Principle 1: Either when building a corpus or when using one, corpus linguists should use research questions in order to engage with their data in a structured and controlled way.

Principle 2: The study of observed language is a means by which language can be investigated and explained.

Principle 3: Corpus linguistics, especially in the form of corpus annotation, is an area where the ontological presuppositions of linguistics become clear.

Principle 4: Because we have a presupposition of reality, through which we can measure hypotheses against observable, relatively objective data, we are able to use the apparatus that has been developed over centuries to support inductive reasoning, including through statistical analysis.

Principle 4': Because we have a presupposition of reality and through that we can measure hypotheses against observable, relatively objective data, we are able to use the apparatus that has been developed over centuries to support inductive and quasi-inductive reasoning, for example through statistical analysis.

Principle 5: By studying corpora, i.e. finite samples of language, we can make general claims about language itself; these claims are probabilistic in nature.

Principle 5': By studying corpora, i.e. finite samples of language bound by space and time, we can make claims claim about language itself, with those claims bound by space and time in the same way as our data.

Principle 6: Corpus linguistics inclines to scientia realis – it is the study of observable language.

Principle 6': Corpus linguistics inclines to scientia realis – it is the study of observable language on which experience may be tested in accordance with Principles 7 and 11.

Principle 6'': Corpus linguistics, drawing on scientia realis, works, as a social science, in a way which is informed by concepts from science – it is the study of observable language on which experience may be tested in accordance with Principles 7 and 11.

Principle 7: Corpus linguistics promotes and is based upon an intersubjectively observable approach to language in which results are repeatable and replicable.

Principle 8: In any given corpus investigation, data and theory are critically considered.

Principle 9: Corpus analysis does not pave a pathway to certainty – it is a problem solving exercise and requires a constant engagement with new problems and the revisiting of old problems.

Principle 10: Corpus linguistics, being closely linked to processes of induction (see Principe 4), also entails the problems of induction which in turn points towards uncertainty (Principle 9).

Principle 10': Corpus linguistics, being closely linked to induction and quasi-induction (see Principe 4), also experiences the problem of induction which in turn points towards uncertainty (Principle 9).

Principle 11: A corpus is of use for testing the falsifiability of a hypothesis about language.

Principle 12: In developing new tools and datasets, corpus linguists are mindful of the need, where possible, to transition hypotheses from the realm of metaphysics to one where important principles of corpus linguistics may be brought to bear, especially Principles 4, 5, 6 and 11.

Principle 13: The observable is not always of importance or utility.

Principle 14: Corpus linguists are critical realists who use corpus data as a way to come into quasi-contact with language in use.

Principle 15: Corpora are theoretically under-determined and hence may support a range of theoretical perspectives.

Principle 16: When building and using corpora, our goal is to facilitate (corpus building) or perform a genuine test (corpus use) and a genuine test must permit the possibility of falsification.

Principle 16': When building and using corpora, our goal is to facilitate (corpus building) and/or perform a genuine test (corpus use) and a genuine test must permit the possibility of falsification while also achieving surprise.

Principle 16'': When building and using corpora, our goal is to facilitate (corpus building) and/or perform a genuine test (corpus use) and a genuine test must permit the possibility of conditional falsification while also achieving surprise.

Principle 17: When conducting a test with corpus data, one must consider the overall impact of any falsification produced by that test.

Principle 17': When conducting a test with corpus data, one must consider the overall impact of any conditional falsification produced by that test. As conditional falsifications accumulate, rejection may be warranted.

Principle 18: Reporting falsifications is arguably more important than reporting corroboration. At the very least, both should be reported.

Principle 18': Reporting falsifications is arguably more important than reporting corroboration. At the very least, both should be reported unless the report is on a previously falsified hypothesis; then neither corroboration or falsification is of value.

Principle 18'': Reporting conditional falsifications is arguably more important than reporting corroboration. At the very least, both should be reported unless the report is on a previously rejected hypothesis – then a falsification of a falsification which led to rejection is of particular interest.

Principle 19: In research practice and debate, corpus linguists should not produce arguments proceeding from dogmatic positions. They are the negation of everything we wish to achieve through an appeal to evidence and objectivity.

Principle 19': In research practice and debate, corpus linguists, as critical rationalists, should not produce arguments proceeding from dogmatic positions. They are the negation of everything we wish to achieve through an appeal to evidence and objectivity.

Principle 20: Corpus linguistics proceeds by convention to use methods, datasets, ontologies etc. accepted by a community of scholars as fit for the investigation of language.

Principle 21: Corpus linguistics proceeds incrementally as testing by falsification proceeds and the cycle of hypothesis formation is driven onwards.

Principle 21': Corpus linguistics proceeds incrementally as testing by falsification proceeds and the cycle of hypothesis formation is driven onwards with the goal of reducing the range of those hypotheses.

Principle 22: Corpus linguists should be open and honest about the limitations of their data and methods.

Principle 23: Corpus data are as provisional as hypotheses and in principle may pass through a similar lifecycle.

Principle 24: Corpora can provide corroborations and falsifications of inverse existential statements. However, they only provide corroboration of existential statements.

Principle 25: Corpus linguists should seek to explore and develop axioms, which are integrated, non-redundant, parsimonious and necessary.

Principle 25': Corpus linguists should seek to explore and develop axioms which are integrated, non-redundant, parsimonious, necessary and as precise as possible.

Principle 26: When using corpora to explore theory, the critical lens provided by Principle 25 should be applied to that theory.

Principle 27: Principles 25 and 26 apply equally to axioms and axioms by convention.

Principle 28: Methods should be combined so as to maximise the potential falsification of any hypothesis explored.

Principle 28': Methods should be combined so as to i.) maximise potential falsifiability; ii.) maximise experimental falsifiers; iii.) maximise the empirical content of the system under examination and iv.) minimize range. All of this is done in the pursuit of simplicity.

Principle 29: Corpus linguists, when selecting data and tools to examine a hypothesis, should select those which maximise the potential falsifiability of that hypothesis while also abiding by other principles, notably Principle 7.

Principle 29': Corpus linguists, when selecting data and tools to examine a hypothesis, should select those which i.) maximise potential falsifiability; ii.) maximise experimental falsifiers; iii.) maximise the empirical content of the system under examination and iv.) minimize range, while also abiding by other principles, notably Principle 7. All of this is done in the pursuit of simplicity.

Principle 30: Consistency is a necessary goal of corpus analysis, annotation and construction.

Principle 31: Intuition is a useful guide to hypothesis formation and revision in corpus linguistics.

Principle 32: Corpus linguistics can be used as the basis for a normative epistemology of language.

Principle 32': Corpus linguistics can be used as the basis for a normative epistemology of language. However, this abstraction away from the individual should not exclude or obscure the individual and variation from the normative.

Principle 33: Corpus linguistics typically works within a 'searchlight' paradigm of research. The needs of theory and axiom construction guide us to specific areas of enquiry.

Principle 34: The weight of any given corroboration that we derive from a study, and thus the confidence we may express in convention, is proportionate to the severity of the test to which we subject our hypotheses.

Principle 35: A corpus represents an amalgam of social and physical interactions.

Principle 35': A corpus, representing socially situated data, is an amalgam of social and physical interactions. Our approach to analysing it should take this into account, recognising the value of the theoretical, the historical and the applied.

Principle 36: Corpora facilitate the study of the institutions and traditions of language.

Principle 37: By applying simplifying assumptions which filter and select corpus data, corpus builders and users may weaken the nature of the quasi-contact with reality established in Principle 14. When doing so, the rationale for the simplification should be presented and its impact on theory and the validity of any tests undertaken considered.

Principle 38: Our approach to probability in corpus linguistics relates to propensity – we believe that there are a large number of forces acting on linguistic propensity, which we need to actively model to the best of our ability, in order to use statistical inferencing meaningfully. The forces form an important context within which language is produced, interpreted and transmitted between generations.

Principle 39:

Where:

U = the set of all forces relevant to the propensity of language

A = the set of forces that the corpus selected allows, on publication, the corpus user to explore

B = the set of forces which, post-publication of the corpus, a corpus user may infer

C = the set of forces available on publication that the corpus user may also seek to derive after publication of the corpus

Wherever possible, i.) corpus builders should maximise the size of A\B; ii.) corpus users should increase the size of B\A; iii.) both corpus builders and users should maximise the size of A∪B as a subset of U.

Principle 40: We believe that those producing language generate output which they believe to be appropriate and well formed from the point of view of their speech community for the purpose of communication in the cognitive, social and physical context in which the utterance was produced, bounded by time.

Principle 41: The evidence for language is the production and reception of language by users and learners of language. The way in which it acts as evidence for them is a way in which it can act as evidence for linguists also.

Principle 42: In situ falsifications regarding beliefs of appropriacy and well-formedness of a linguistic feature or structure based on Principle 40 may occur where appropriacy and well-formedness is rejected by a person in the intended audience of the linguistics production in question.

Principle 42': In situ falsifications regarding beliefs of appropriacy and well-formedness of a linguistic feature or structure based on Principle 40 may occur where appropriacy and well-formedness is rejected by a person in the intended audience of the linguistic production in question or the producer of the language.

Principle 43: Principle 40 leads to reflexive behaviour on the behalf of a recipient of a linguistic production (hearer) – they choose to accept a linguistic production, reject it or, potentially, to modify

their own view of what is appropriate based on the communicative event in question. Accordingly, the corpus linguist may come into quasi-contact with such reflexive events also.

Principle 44: Researchers should make their best endeavours to ensure that the research results they publish will remain repeatable. This entails archiving versions of corpora and tools used in such studies, describing clearly the goals and methods of the study and ensuring that analytical abstractions on which a study is based (e.g. linguistic analyses, statistical analyses etc.) are well described in such a way as to promote repeatability.

Principle 45: Version management for corpora and tools should be considered at the start of a project and then maintained over time.

Principle 46: Sampling structures in corpus linguistics are seeking to divide language into chunks within which it may be expected that language varies, i.e. the sampling structure is composed of nominal forms within which language variation occurs, typically genres, registers or modes of communication. They embody ontological knowledge.

Principle 47: The process of investigating corpus data has helped linguists show, by repeated investigation, how elements of ontological knowledge can, by convention, be assumed to be meaningful when used to structure our observation of language through sampling. This, in turn, controls our view of the process of replication enshrined in Principe 48.

Principle 48: In undertaking repetition and replication, corpus linguists should be clear about what they are doing, and their goals in doing so. Treatment design and experimental design should be clearly identified and, in the case of replication, a clear explanation given of what surprise has been introduced, what hypotheses are potentially impacted by this, what potential outcomes were anticipated and what results were achieved.