School of Linguistics & Language Studies
APPLIED LINGUISTICS & DISCOURSE STUDIES

ALDS 3701: Corpus Linguistics   (Fall 2011)

Instructor: Dr. Guillaume Gentil (Associate Professor)
Email: guillaume_gentil@carleton.ca
Website: carleton.ca/slals/people/gentil-guillaume
Class Hours: Mondays and Wednesdays, 6:00 - 7:25 p.m.

Course Description:

The computer-assisted analysis of language samples or corpora has become an important methodology in language studies with significant implications for language description, discourse analysis, technical writing, translation, and language teaching. With a hands-on approach, this course will introduce you to corpus linguistics, familiarize you with corpus resources and corpus-analysis software, and show you how to design, build, and use your own corpora. You will thus learn about resources and techniques employed by professional writers, translators, language teachers, dictionary makers, language material developers, and other language specialists interested in revealing, quantifying, and exploiting patterns of language use that only the power of computers can reveal.

Depending on participants' interests, applications will be explored in such areas as language variation, grammar, lexicography, phraseology, technical writing, translation, learner language, the study of ideology and culture, critical discourse analysis, stylistics, and language teaching.

Although this course will introduce you to basic statistical and computing techniques for language and discourse analysis, it does *not* require knowledge of computer programming. It is *not* a course in computational linguistics or natural language processing, but it does make use of resources developed by computational linguists. Several classes will take place in a computer lab. Evaluation will consist of a portfolio of in-class activities and a small corpus-based research project.

Course Objectives

- To identify the main types of corpora (reference/specialized, monolingual/multilingual, spoken/written, comparable/parallel, learner corpus, monitor corpus, diachronic/synchronic corpus, untagged/tagged/parsed)
- To identify the primary corpora available online
- To introduce key corpus linguistics (CL) tools and concepts (including frequency, dispersion, concordances, collocation, colligation, cluster, keyness, semantic preference/prosody, type/token, hapax, lemma/word form, open-choice vs. idiom principle)
- To raise awareness of issues in corpus design (representativeness, balance, and sampling)
- To introduce principles and techniques in corpus compilation (such as data capture, data encoding, data markup/annotation, part-of-speech tagging)
- To familiarize students with corpus linguistics software, online corpus query processors (CQP), and search techniques (such as regular expressions)
- To introduce basic descriptive and inferential statistics used in corpus linguistics (including chi-square test, log-likelihood test, MI-score, t-score, z-score)
- To explore, illustrate, and practice selected applications in corpus linguistics (e.g., dictionaries and grammars, writing and translation aids, discourse analyses of ideology and culture, studies of language acquisition and language variation, data-driven language learning and language teaching)
- To explore links between corpus linguistics and related fields such as applied linguistics, terminology, lexicography, phraseology, stylistics, critical discourse analysis, and natural language processing
- To show how to design, build, and use small specialized corpora and how to assist in the design and reporting of a small corpus-based study