第 6 回　11 月 14 日（火）

「日本語：数億語のコーパスを作って調べるとみえてくる頻出語、頻出表現」

講師：望月 源　　東京外国語大学准教授

我々が生活の中で実際に使っている言語を「データ」として集めて電子的に利用可能にすると「コーパス」になります。今日では、WWW のホームページ、ブログ、電子ニュースなど、言語データとして利用できる情報が大量に流通しています。本講義では、実際にどのようにしてコーパスを作成するのか、いくつかの実例を示して簡単に説明します。また、日本語の数億語規模のコーパスを調べて、どんな語、どんな表現が実際によく使われているのかを確かめます。実際のデータから得られる結果は、皆さんの直感と合うか、あるいは予想外になるのか、理由も含めて一緒に考えたいと思います。