はじめに

ことばのデータサイエンス

情報化時代といわれる現代において，**データマイニング**の技術が大きな関心を集めています。データマイニングは，情報学や統計学の技術を用いて，大規模なデータから有益な知識を効率よく発見する理論および技術の総称です[*1)]。

2010年代に入ってからは，**ビッグデータ**や**データサイエンス**という用語が頻繁に使われるようになり，学術書や研究論文のみならず，一般書やビジネス雑誌でそれらの用語を見かけることも多くなってきました。このような流れの中で，**テキストマイニング**という技術も大きな注目を集めるようになりました。

テキストマイニングは，**テキストアナリティクス**とも呼ばれ，テキスト（言語データ）を対象とするデータマイニングのことです。テキストマイニングは，大量の言語データを解析し，データの背後に潜む有益な情報を探し出すことを主な目的としています。具体的には，テキストにおけるキーワードの抽出，特定のキーワードと共起する語句の特定，テキストの自動分類などに活用されることが多いです。

そして，「ことばのデータサイエンス」という本書のタイトルは，テキストマイニングやテキストアナリティクスのように言語を対象とするデータサイエンスという意味を持っています。

[*1)] データマイニングの「マイニング（mining）」は，「掘り出す」という意味を持つ“mine”という動詞の現在分詞です。つまり，データマイニングは，鉱山から鉱石を掘り出すように，データの山から何らかの知見を掘り出すイメージです。

本書の特色

本書は，計量的な言語研究の入門書です。具体的には，コンピュータや統計学を用いた言語研究の方法を解説し，実際の言語データの解析事例を多く紹介します。しかし，言語学や文学の研究者以外の方々，一般の方々にも読んで頂けるように，できるだけわかりやすい分析事例を取り上げ，専門用語などには註釈をつけています。また，主に文系の読者を想定し，統計処理の方法を解説する部分で，四則演算（足し算，引き算，掛け算，割り算）などによる簡単な計算式を超える内容に関しては，イメージ図や言葉で説明しています[*2)]。

なお，本書は，特定のソフトウェアのハウツー本ではありません。したがって，本書で紹介されている処理を読者が実際に行う際は，別途ソフトウェアのマニュアルや参考書を参照する必要があります[*3)]。ただ，読者の便宜を考慮し，本文や註釈などで，本書執筆時点でおすすめのソフトウェアなどを積極的に紹介しています[*4)]。

本書をきっかけに，計量的な言語研究に興味を持って頂ければ幸いです。

2019 年 8 月

小林　雄一郎

[*2)] 原則として，第 7 章までは計算式を示していますが，多変量解析を扱う第 8 章以降ではイメージ図と言葉による説明が多くなっています。

[*3)] 本書を特定のソフトウェアのハウツー本にしなかったのは，特定のソフトウェアに依存した研究では，そのソフトウェアの限界が研究そのものの限界となり得るからです。また，ハウツー本はすでに多く出版されている上，操作方法などの説明が数年で古くなってしまう場合もあるからです。

[*4)] 紹介しているソフトウェアの中には，最新の OS に対応していないものや，特定の OS のみに対応しているものがあります。また，ウェブサイトの URL やスクリーンキャプチャは本書執筆時点のものであり，今後変更されることもあり得ます。