コーパス基盤の言語研究

李　　在鎬

１．コーパスに基づく言語研究

コーパスを利用した言語研究、概してコーパス言語学（corpus linguistics）は数理言語学の一分野であるとされている。コーパスとは何か、コーパス言語学とは何かということをめぐっては諸説あるが、概念的な定義の問題には深入りせず、方法論的特徴に焦点を当て考察する。

コーパスとは、元々「言語分析のための文字言語・音声言語の資料の集合体」であるとされているが、近年では電子媒体の資料を指すことが多く、電子コーパス（electronic corpus）とほぼ同義のものと位置づけられている。

コーパスを使った言語研究の特徴として、４点が広く知られている（Leech 1992）。

１．言語能力より言語運用に中心をおく。
２．言語の普遍的特徴の解明より、個別言語の記述に中心をおく。
３．質的な言語モデルのみならず、数量的な言語モデルにも中心をおく。
４．言語研究における合理主義的な立場よりも、経験主義的立場に立つ。

コーパス言語学では言語行動の集積であるコーパスデータを忠実に分析することを目的としているため、言語運用に中心をおくアプローチといえる。もっと正確には、コーパス言語学的見方では、言語運用と言語能力は表裏一体のものとして捉えられ、言語運用から言語にアプローチする[1]。この１の特徴から２の方向性が導かれ、言語現象の個別的特徴に注目する。そして、分析手法として、３の数量的・計量的手法が用いられる。コーパス言語学では反復される言語的事象は重要であると考えられており、一般的で典型的なものを記述する。こうした特徴を持っていることから、コーパス言語学は、経験基盤主義に基づく言語研究の方法論であるとされている。

次に、コーパスを用いる積極的な意義について考えてみよう。二点としてまとめられる（Biber, Conrad and Reppen 1998）。

１．具体的な言語の使用実態をパターン化し、調べることができる。
２．網羅的に現象を収集することができ、データの偏りが解消できる。

経験科学としての言語研究の意義を考えてみた場合、思弁に基づく理論の精緻化ではなく、具体的な言語使用を分析し、その動機づけを明らかにすることがより重要である。コーパスに基づく言語研究は大量のデータを組織化することで、言語使用の実態をパターン化し、可視化することができる。また、言語の研究にとって、データの重要性は強調するまでもないことである。作例基盤の研究を否定するわけではないが、人間のイマジネーションには限界があり、必然的に偏ったデータ収集がなされてしまう。しかし、コーパスを使用した場合、入手可能な全用例を解析したり、無作為抽出を行うことによって、扱うデータの偏り

を回避することができる。これらの利点を生かすことで、従来の研究では、見過ごされてきた現象を発見できる可能性が開かれる。特にコーパスによる語彙研究の手法は、その生起文脈を客観的に記述することができることから言語研究の単なる手法の域を超え、従来の言語研究では認識されてこなかった問題や現象を発掘し、解明していく新たな言語研究のパラダイムとして確立しつつある（cf. Tognini-Bonelli 2001）。

2．良いコーパス調査のために

コーパス分析による調査および研究を行う際、留意すべき点をまとめる。もっとも基本的なこととして、研究全体に対する見通しを持って調査を望むことが重要である。どういう目的で、どのようなコーパスを利用し、どのような調査を行うのか、その結果に対してどのような予測や仮説を持っているかといった基本事項に対し、明確な方向性を持って、調査を臨むべきである。

さて、コーパス言語学に関連する留意事項を検討する上で、計量言語学の知見は役立つ。とりわけ良い調査と言えるための条件として、次の5点が指摘されてきた。

1．妥当性：まさに知ろうとする目的を正しくつく調査方法であること
2．信頼性：同じ対象に同じ操作を加えた結果があまり大きく変動するようなものでないこと
3．客観性：結果が主観によってまちまちになるような操作ではないこと
4．再現性：調べる対象の実際の姿が操作を加えた結果から正しく見極められること
5．適応性：その操作が実際に無理なく行え、かつ妥当な結果をもたらすようなものであること

これらは標準抽出で代表される計量的研究において主に問題になることであるが、データの扱い方に対する汎用的条件になるので、良いコーパス調査の条件として理解することもできる。具体的には、1として手法の適切性の問題がある。1の問題がクリアされないと、調査そのものが無意味なものになることも少なくない。例えば、日本語の格助詞の分布を単純に文字列のみで数えるような調査方法では妥当な結果が得られない。また、2として、同じコーパスからデータを抽出する場合、抽出の量が少なければ偶然性が高くなり、反対に多ければ分析者によって結果が違ってしまうという人為的なミスが生じやすくなる。したがって、用例の数が極端に少なすぎても、多すぎても信頼性を損ねてしまう可能性がある。最適な用例数を見つけるためには、事前調査を行うなり、関連研究を見るなどして、ターゲットとなる現象をある程度見極めておく必要がある。そして、3として、調査の単位が途中で変わってしまうようでは客観的な結果が得られない。4として、抽出データが母集団の偏った一部のみを取り出したものでは、母集団の全体性を正しく再現したものではない。最後に、5として、分析者が調査に対する十分な理解がなく、綿密なデザインがないまま調査を行った場合、調査の途中で頓挫したり、間違いを含んだデータを生み出す可能性が高くなる。

コーパス調査は大量の言語データと根気強く付き合っていく態度で臨むべき作業である。したがって、実際の調査の前に調査の方針や基準などを明確にし、調査の事前作業として先行研究を確認する作業や複数の人で企画案を粘り強く練っていくという作業が不可欠である。実際の調査では様々な制約から上述のすべての要件を満たすことは困難な場合が多いが、その場合、もっとも重視すべきは、1の条件である。

3．最後に

本稿ではコーパスを用いた研究の意義と良い調査を行うために心がけてほしい事項について述べた。紙幅の都合上、実際のコーパスデータや分析ツールの詳細までは述べることができなかったが、http://www30.atwiki.jp/corpus-lingにおいて、詳細を提示しているので、合わせて参照していただきたい。

注

1　Sinclair（1991:104）などでは、合理主義言語モデルが主張する二分法（Saussureの言語論であれば、ラングとパロールであり、Chomskyの言語論であれば行動と精神、言語運用と言語能力、あるいはE言語とI言語の区分）は、分析者の誤った思考によるもので、同一の現象を異なった見方で捉えているだけであると主張する。

参考文献

Biber, D., S. Conrad and R. Reppen (1998) *Corpus Linguistics: investigating language structure and use.* Cambridge: Cambridge University Press.

伊藤雅光（2002）『計量言語学入門』大修館書店

Leech, G. (1992), Corpora and theories of linguistic performance: in Svartvik, J. ed., *Directions in corpus linguistics: proceedings of Nobel symposium* 82, Berlin and New York, Mouton de Gruyter, 125-148.

中本敬子・李在鎬・黒田航（編著）（2010印刷中）『新しい認知言語学研究法入門』ひつじ書房

Sinclair, John. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work.* Amsterdam: John Benjamins.