С. Шітанда,
Львівський національний університет
імені Івана Франка

コーパス言語学観点における現代日本語書きことば語彙の変遷過程解明のための予備考察
(Пропедевтика для пояснення перехідного процесу між словниковим складом та граматикою у сучасній японській письмовій мові з точки зору корпусної лінгвістики)

Національний інститут японської мови та лінгвістики (NINJAL) розробив корпус «Таййо» для того, щоб дослідити процес формування та становлення сучасної японської мови на початку 20 століття. Метою цієї роботи є виконання обчислювального лексикологічного аналізу та рішення для процесу переходу лексики в період становлення сучасної японської писемності на початку 20 століття, використовуючи корпуси «Таййо» і KOTONOHA.

Ключові слова: *корпус, обчислювально-лексикологічний аналіз, сучасна японська писемність*

Национальный институт японского языка и лингвистики (NINJAL) разработал корпус «Таййо» для того, чтобы исследовать процесс формирования и становления современного японского языка в начале 20 века. Целью этой работы является выполнение вычислительного лексикологического анализа и решения для процесса перехода лексики в период становления современной японской письменности в начале 20 века, используя корпуса «Таййо» и KOTONOHA.

Ключевые слова: *корпус, вычислительно-лексикологический анализ, современная японская письменность*

National Institute for Japanese Language and Linguistics (NINJAL) produced the "Taiyo" corpus in order to research the process of formation and establishment of contemporary Japanese language in the early 20th century. The purpose of this thesis is to perform computational lexicological analysis and solution for the transition process of vocabulary in the period of the establishment of contemporary Japanese written language of the early 20th century, using "Taiyo" and KOTONOHA corpora. In the thesis will be discussed: 1) the purpose, theory and methods of computational lexicology, 2) the results of the case analysis of vocabulary of contemporary Japanese written language based on "Taiyo" and KOTONOHA corpora.

Key words: *corpus, computational lexicological analysis, contemporary Japanese written language, Taiyo, KOTONOHA*

0. 本稿の研究対象と目的

本稿の目的は、複数のコーパスを用いた、20世紀初頭の現代日本語書きことば確立期20世紀末から21世紀初頭の現代日本語書きことばの語彙の変遷過程の計

量語彙論的分析と解明のための理論的・実践的予備考察である。現代日本語の書きことばは、明治後期から大正時代にかけての20世紀初頭に確立したが、この時期に多くの近代的概念を表す語彙と、口語と文語の距離を近づける言文一致の口語文が広まった。[国立国語研究所 2005a, 3]方法論的には、複数のコーパスを応用して、日本語の量的語彙構造史構築のためのモデル分析を試みる。本論では、1)先行研究における計量語彙論の理論・方法、2)複数のコーパスに基づく書きことば語彙の範例的分析と結果、について論じる。計量語彙論的分析の際、量的語彙構造の比較分析の対象となるのは、以下のコーパスである。1)中古語彙：宮島達夫他編『古典対照語い表』[宮島 1989]。上代から中世までの古典14作品の語彙の自立語を統計データとする。2)「太陽」コーパスの現代語書きことば確立期語彙：国立国語研究所編(2005)『雑誌「太陽」日本語データベース』。対象範囲は、雑誌「太陽」の1895、1901、1909、1917、1925年に刊行された計60冊、総記事数3408、合計約1445万字である。3)現代書き言葉語彙：「ことのは」の『現代日本語書き言葉均衡コーパス』 (BCCWJ)。1986年から2005年までの生産実態（出版）サブコーパス、流通実態（図書館）サブコーパス、非母集団（特定目的）サブコーパス からなる一億語を含む現代書きことば語彙コーパスで、国立国語研究所によるオンライン公開A(少納言)とオンライン公開B(中納言)がある。(http://www.ninjal.ac.jp/kotonoha/ex_6.html)

1. 先行研究における量的語彙構造分析法の可能性と問題点

先行研究における計量語彙論、及び量的語彙構造を解明する分析方法について言及する。林(1968,1971)は「基幹語彙選定法」を提案したが、それは「語彙素の使用度数と使用範囲から基幹語彙を選定する方法」[伊藤 2009, 175]で、「使用度数が高く、使用範囲が広い語彙素ほど基本度が高い。」[伊藤 2009, 175] 基本度に基づき語彙区分を行い、対象は新聞3誌1年分(1966)とし、記事のテーマ分野（ジャンル別）、階層12区分から成る「共時的使用範囲」を規定する。その際、高頻度・広範囲の語彙に限定される。これに対し、伊藤(2008)「量的語彙構造分析法」は、林の選定法を構造史の分析に応用し、「各時代の語彙を、基本度の観点から小語群に区分」[伊藤 20 09, 175]し、語彙構造を解明する。この方法の「通時的使用範囲」の内容は、上代、中古、中世の三時代区分である。通時的使用範囲の総延べ数は、各時代の使用度数であり、分析表は各時代別に作成される。伊藤の分析法では「ある語彙素の各時代の使用度数と通時的使用範囲を基準にして、その時代における語彙素の位置づけが決定される」[伊藤 2009, 176]。分析対象の範囲としては、基幹（基本）語彙だけでなく、特殊語彙も対象となり、低頻度かつ狭範囲の語彙も含めた語彙全体を対象とする。語彙の量的構造史は、抽象度によって2つのレベル（二次元及び一次元モデル）に分けられる。二次元モデルは、語彙範囲と語彙頻度に基づき、「表層構造」とされる。語彙の基本度の抽象度に基づく一次元モデルを、深層構造と規定している。伊藤は、一次元モデルの深層構造に、「量的語彙構造の法則性」（高頻度かつ広範囲の語彙素で構成される語彙は、基本度が高く（特徴度が低く）、低頻度かつ狭範囲の語彙素で構成される語彙は特徴度が高い（基本度が低い））を認めている。[伊藤 2009, 196]各時代の量的語彙体系モデルを時系列順に並べ、「語彙の量的体系史の継時態モデル」を構築する。

レベル / 時代		
レベル**1** 語彙体系・二次元モデル 具象的レベル	現代語彙 現代の量的語彙体系・　二次元モデル	近世語彙 近世の量的語彙体系・　二次元モデル
レベル**2** 語彙構造・二次元モデル 表層構造	現代の量的語彙構造・　共時態二次元モデル	近世の量的語彙構造・ 共時態二次元モデル
レベル**3** 語彙構造・一次元モデル 深層構造	現代の量的語彙構造・　共時態一次元モデル**1-2**	近世の量的語彙構造・ 共時態一次元モデル**1-2**

上図は、上代から中世に及ぶ伊藤の量的語彙構造史の継時態・多層構造システムモデルを近世語・現代語に転用・拡大した、筆者によるモデルである。伊藤の「量的語彙構造分析法」の問題点は、1)語彙データが『古典対照語い表』のみであるため、語彙のジャンルが13の古典文学作品に限定されており、他分野の語彙データが欠けている、2)量的語彙構造は、「語彙素(Lexeme)の量的分布状態」を示すため、語場(Semantic　field)を示す質的語彙構造のように、意味論的関連・変化を捉えられない、3)レベル2の表層構造とレベル3の深層構造の差が大きく、「量的語彙構造の法則性」だけが語彙構造の深層構造を構成要因であるとは言えない、の3点に集約される。

2. 複数のコーパスに基づく現代書きことば語彙の範例的分析

今回の分析では、伊藤の「量的語彙構造分析法」(2008)を応用し、「太陽」コーパス　(国立国語研究所編(2005)『雑誌「太陽」日本語データベース』) と KOTONOHA「現代日本語書き言葉均衡コーパス」の「少納言」(http://　www.kotonoha.gr.jp/shonagon/)に限定した。表層構造のうち、現代語の前身の「汎時代語」を検索し、特に頻度差（基本度と特徴度を測定する）を検証した。対象は以下の四語彙：1)「こと（事）」（上代中古中世の三時代共通語彙、汎時代語、高頻度）、2)「このごろ」（上代中古中世の三時代共通語彙、汎時代語、中頻度）、3)「あさひ」(上代中古中世の三時代共通語彙、汎時代語、低頻度a)、4)「いはゆる（所謂）」(上代中古中世の三時代共通語彙、汎時代語、低頻度b)とし、以下の分析結果を得た。

検索語 / 発行年	1895	1901	1909	1917	1925	合計	平均値頻度
こと	**7794**	**8630**	**8516**	**8599**	**8974**	**42513**	**0,002942**
事	**11305**	**10289**	**9163**	**8217**	**6361**	**45335**	**0,003137**

このごろ	**5**	**2**	**0**	**0**	**3**	**10**	**6,919e-7**
この頃	**9**	**1**	**23**	**27**	**40**	**100**	**6,919e-6**
此の頃	**5**	**6**	**14**	**9**	**15**	**49**	**3,391e-6**
あさひ	**0**	**0**	**0**	**0**	**0**	**0**	**0**
朝日	**41**	**54**	**28**	**22**	**27**	**172**	**1,190e-5**
いはゆる	**12**	**4**	**0**	**0**	**7**	**23**	**1,592e-6**
所謂	**553**	**694**	**817**	**472**	**381**	**2917**	**2,018e-4**

範例的分析結果**(1)** 太陽コーパス **(1895,1901,1909,1917,1925)**

検索語/ジャンル	書籍 71-05 6230万語	雑誌2001-2005 440万語	新聞2001-2005 140万語	合計	
こと	297899	**22748 (0,00517)**	**5414 (0,00387)**	**28162**	**0,004856**
事	101228	**8288 (0,00188)**	**5100 (0,00364)**	**13388**	**0,002308**
このごろ	198	**17 (3,86e-6)**	**5 (3,57e-6)**	**22**	**3,79e-6**
この頃	518	**43 (9,77e-6)**	**1 (7,14e-7)**	**44**	**7,59e-6**
此の頃	7	**1 (2,27e-7)**	**0 (0)**	**1**	**1,72e-7**
あさひ	114	**6 (1,36e-6)**	**2 (1,43e-6)**	**8**	**1,38e-6**
朝日	998	**189 (4,29e-5)**	**143 (1,02e-4)**	**332**	**5,72e-5**
いはゆる	14	**1 (2,27e-7)**	**0 (0)**	**1**	**1,72e-7**
いわゆる	3108	**201 (4,57e-5)**	**36 (2,57e-5)**	**237**	**4,09e-5**
所謂	140	**3 (6,82e-7)**	**0 (0)**	**3**	**5,17e-7**

範例的分析結果**(2)** 現代日本語書き言葉均衡コーパス「少納言」**(2001-05)**

3. 終わりに **–** 範例的分析の結果と今後の課題 **-**

汎時代語「こと（事）」は、両コーパスで頻度が高く、常に高い基本度を示すのに対し、中頻度語「このごろ」、低頻度語「あさひ」は、基本的には中頻度以下の範疇に留まっている。低頻度a)語「朝日」が比較的頻度が高いのは、現代語で固有名詞としての用法が多いためと考えられる。「所謂」は低頻度語でありながら、20世紀初頭では論理的文章に特有の表現として多用され、21世紀に入る

と再び減少傾向に転じる。また、これらの語彙がひらがな表記か漢字表記かで、コーパス間で出現度に大きな変化が見られる。[国立国語研究所 2005b, 157] 平均値頻度では、ひらがな表記「こと」は1917年以降、漢字表記「事」を上回っている。今後の課題として、上代・中古・中世の三時代共通語彙のうちの汎時代語（130語）について、両コーパスで検証し、近代を除く四時代共通語彙を抽出し、現代語の基本語彙集の基本語彙との比較検証を行い、量的語彙体系の核となる通時的・継時的基本語彙の形成が求められる。さらに現代語彙の量的構造体系の二次元モデル、量的語彙構造・共時態二次元モデル・一次元モデルを構築する必要がある。

参考文献

1. 安部清哉、斉藤倫明、岡島昭浩、半沢幹一、伊藤雅光、前田富祺(2009)『シリーズ日本語史2 語彙史』岩波書店

2. 国立国語研究所編(2005a)『太陽コーパス　雑誌「太陽」日本語データベース』 国立国語研究所資料集 15　博文館新社

3. 国立国語研究所編(2005b)『雑誌『太陽』による確立期現代語の研究　『太陽コーパス』研究論文集』国立国語研究所報告　博文館新社

4. 宮島達夫、中野洋氏、鈴木泰氏、石井久雄(1989)『フロッピー版古典対照語い表および使用法』笠間書院

5. http://www.ninjal.ac.jp/kotonoha/ex_6.html