言語機能研究のための日本語の単語・文字データベース

近藤　公久、　天野　成昭
NTT 基礎研究所

【はじめに】人間の言語機能の解明研究において、実験に用いる言語刺激の特性を統制する必要がある。なぜならば、言語の特性の違いが実験の結果に影響を及ぼすことが知られているからである。しかし、この統制は非常に困難な作業であった。それは、統制しなければならない特性に関する十分なデータが存在しなかったからである。そこで、我々は、言語認知過程において基本的な単位である単語と文字に関する特性値データベースを構築した。

【特性値データベース】1）2）今回構築したデータベースに収録されている単語および文字の特性値データについての概略を示す。

2.1　単語親密度評定値 3）4）収録単語：新明解国語辞典 5）の見出し語および小見出し語として収録されている単語のうち、自立語約 80，000 語。収録データ：30 才以下の男女 32 名による「なじみの程度」の 7 段階評定の平均値。文字提示、音声提示、及び文字・音声同時提示の 3 条件における評定値を収録。

2.2　単語の表記の妥当性評定値 6）　収録単語：単語親密度に同じ。収録データ：30 才以下の男女 20 名による、各単語に対する、漢字かな混じり表記（新明解国語辞典の見出し表記と「とも書く」で示された表記）とひらがな・カタカナ表記の「もっともらしさ」の 6 段階評定の平均値。

2.3　単語のアクセントの妥当性評定値 7）収録単語：単語親密度に同じ。収録データ：両親、本人ともに東京生まれ東京育ちの 20 代の男女合わせ 10 名による、各単語の各アクセント型に対する「もっともらしさ」の 5 段階評定の平均値。それぞれの単語に対して評定対象としたアクセント型は、新明解国語辞典 5）に示されているアクセント型と一部評定者により付加されたアクセント型。

2.4　単語出現頻度 8）出現頻度調査対象：朝日新聞朝夕刊 12 年分(1985－1996)の見出し、図表を除くすべての記事。収録データ：対象とした全記事中に各単語が出現した回数。記事文からの単語抽出にはＮＴＴ開発の形態素解析システム「すもも」9）を使用。１回以上出現した単語は約 16 万語。

2.5　文字親密度評定値 10）11）収録文字：情報交換用漢字符号 12）にある 6,879 文字から罫線用素片 42 文字を除いた 6,837 文字。収録データ：単語親密度と同じ条件の 24 名による各文字に対する「なじみ」の程度の 7 段階評定の平均値。

2.6　文字の複雑度評定値 18）収録文字：文字親密度に同じ。収録データ：単語親密度と同じ条件の 24 名による各文字に対する主観的な複雑さの 7 段階評定の平均値。

2.7　漢字の読みの妥当性評定値 14）　収録文字：情報交換用漢字符号 12）に存在する漢字 6,355 文字。収録データ：30 才以下の男女 24 名による各漢字に対する読み方の「もっともらしさ」の 7 段階評定の平均値。それぞれの漢字に対して評定対象とした読み方は、新明解漢和辞典 15）に示されているすべての読み方と単語親密度の評定を行った単語約 80,000 語中に出現する読み方。

2.8　文字出現頻度 16）　頻度調査対象：単語出現頻度に同じ。収録データ：対象とした全新聞記事中に各文字が出現した回数。

【まとめ】日本語の単語と文字に関する特性値データベースを構築した。我々は、本データベースに収録されている特性値と様々な言語処理過程との関係にいても既に解

析を行っている 6）14）17)。また、親密度を用いた言語能力テストを開発し、その有効性
も確認した 18)。今後、本データベースを書籍（ＣＤ－ＲＯＭ付）として出版する予定である。

文献

1）Kondo,T.,Amano,S.,& Maxuka,R.(1996,March) “ Japanese lexicon database for psycholinguistic research. ” Poster session presented at the CUNY Conference on Human Sentence Processing, New York, NY.
2）近藤，天野（1997）“ 認知科学研究のための日本語の特性に対する主観的評定値データベース ”，日本認知科学会第 14 回大会講演論文集.72-73
3）Amano, S.,Kondo, t.,& Kakehi, K.(1995) “ Modality dependency of familiarity ratings of Japanese words. ” Perception ＆ Psychophysics 57, 598-603.
4）天野，近藤（1998）“ 言語心理学のための日本語単語親密度データベース ”，音響学会春季講演論文集Ｉ.
5）金田一他編（1994）新明解国語辞典第四版，三省堂，東京.
6）Ainsworth-Darnell, K.& Kondo, T.(1998, Janualy) “ Beyond orthographic depth: Similarities in the processing of words in Kanji and Hiragana. ” Paper presented at Annual Meeting of Linguistic Society of America, New York, NY.
7）天野，近藤（1995）“ 日本語単語アクセントの大規模評定実験 ” 音響学会秋季講演論文集Ｉ.
8）近藤，天野（1998）“ 新聞中の日本語単語の出現頻度と単語親密度の関係 ”，日本認知科学会第 15 回大会講演論文集.
9）鷲坂,山崎,廣津,尾内（1997）“ 情報検索のための高速日本語形態素解析シス　テム「すもも」” 情報処理学会全国大会.
10）近藤，天野（1995）“ 日本語の文字に対する親密度 ”，日本心理学会第 59 回大会発表論文集.
11）Kondo, T. & Amano, S.(1996, August) “ Familiarity ratings of Japanese Kanji characters. ” Poster session presented at the International Congress of Psychology, Montreal, Canada.
12）日本工業標準調査会 “ 情報交換用漢字符号（JIS X O208－1990）”,（日本規格協会,東京，1990）.
13）近藤，天野（1997）“ 漢字の親密度と複雑度 ”，日本心理学会第 61 回大会発表論文集.
14）Kondo, T. & Wydell, T.(1997, December) “ Nature of naming latency for Japanese Kapji words. ” Poster session presented at the International Conference on Again Language Processing, Nagoya, Japan.
15）長澤他編（1990）新明解漢和辞典第四版，三省堂，東京.
16）近藤，天野，横山，野崎（1996）“ 漢字の親密度と出現頻度の相関 ”，日本心理学会第 60 回大会発表論文集.
17）鈴木他（1998）“ 親密度を統制した単語了解度試験における反応傾向 ”，音響学会聴覚研究会資料 H-98-47.
18）Kondo, T., Wydell, T., & Amano, S. (1998, March) “ Language processing and reading skills in Japanese. ” Postersession presented at the CUNY Conference on Human Sentence Processing, New Brunswick, NJ.