「なたね」説明書

2021 年 7 月　仁科喜久子（執筆）

言語資源の名称　日本語学習者作文コーパス「なたね」

言語資源の説明

1. 日本語学習者作文コーパス「なたね」の概要

日本語学習者作文コーパス「なたね」は、日本語学習者から収集した作文に対して 2007 年から 2011 年にかけて 3 名の日本語教師によって添削を行い、誤用タグを付与した学習者作文コーパスである。学習者 192 人による 285 件の作文は東京工業大学、中国・西安交通大学で収集したものおよび「インド・プネー市学習者コーパス」を収録している。作文テキストに対し、誤用タグの大分類として誤用の対象（3,886 件）、内容（3,667 件）、要因・背景（1,470 件）、合わせて 9,023 件のアノーテーションを XML 形式で提供している。2021 年現在、本資源の検索インターフェースは

https://hinoki-project.org/

において学習支援作文検索「なたね」として稼働している。

作文データを公開するに当たっては、関係者を通じて、学習者一人ひとりの許諾を得ている。なお、本コーパスは、2010-2012 年度日本学術振興会科学研究費研究「日本語学習者誤用コーパスを利用した作文システムの開発」（挑戦的萌芽研究：22652048）の助成を受けて作成されたものである。

なお「なたね」は「ひのきプロジェクト」（メンバー：阿辺川武、八木豊、ホドシチェクボル、仁科喜久子）の一部であり、語と語の共起検索システム「なつめ」、学術論文作成支援「ナツメグ」という日本語学習作文支援システムとともに提供している。

以下、作文データの収集方法、配布データの概要、誤用タグの策定と付与、誤用の種別、誤用の要因、誤用の内容、統計データ（母語別作文数、母語別学習者数、誤用対象数、誤用の内容、誤用の要因・背景）の順で詳細を述べる。

２．作文データの収集方法

作文データを収集する際には西安交通大学 曹紅荃教授、「インド・プネー市学習者コーパス」については国立国語研究所 Prashant Pardeshi 教授の協力を得た。作文は日本語科目の授業中の課題作文、実験のために与えられた課題作文をまとめたものであるため、作文タイトルもさまざまである。たとえば、「将来したいこと」（初級者）「自己紹介」「日本人がおかしいとおもったこと」「インターネット」「高齢化社会」「男女平等」「研究室見学」など初級者から上級者までによって、様々な内容で書かれている。これらの作文データを分析し、誤用タグを付与する作業は東京都立大学黒田史彦准教授、前出の曹紅荃教授、当時東京工業大学教授仁科喜久子の 3 名を中心に 2008 年から行った。開始に当っては、「誤

用タグ策定」のための議論を前出の 3 名および八木豊氏（㈱ピコラボ）が参加し、文章全体を多層的に分析する方法を考えた上で、日本語教育専門家 3 名によるタグ付けを開始した。当初は同じ Excel 表上に 3 名がそれぞれの判定を書き込んだ後、異なる場合は意見の調整をし、統一化した。その後、大量データ作業のための誤用タグ付与における揺れを防ぎ、客観性を保つため、アノテーションツールとして Slate を使用することとし、製作者である東京工業大学 徳永健伸教授、同研究室大学院生 Dain Kaplan 氏に加わっていただいた。

３．誤用タグの策定と付与

図 1 は、日本語教育の専門家 3 名が策定したタグ付け項目最終版である。

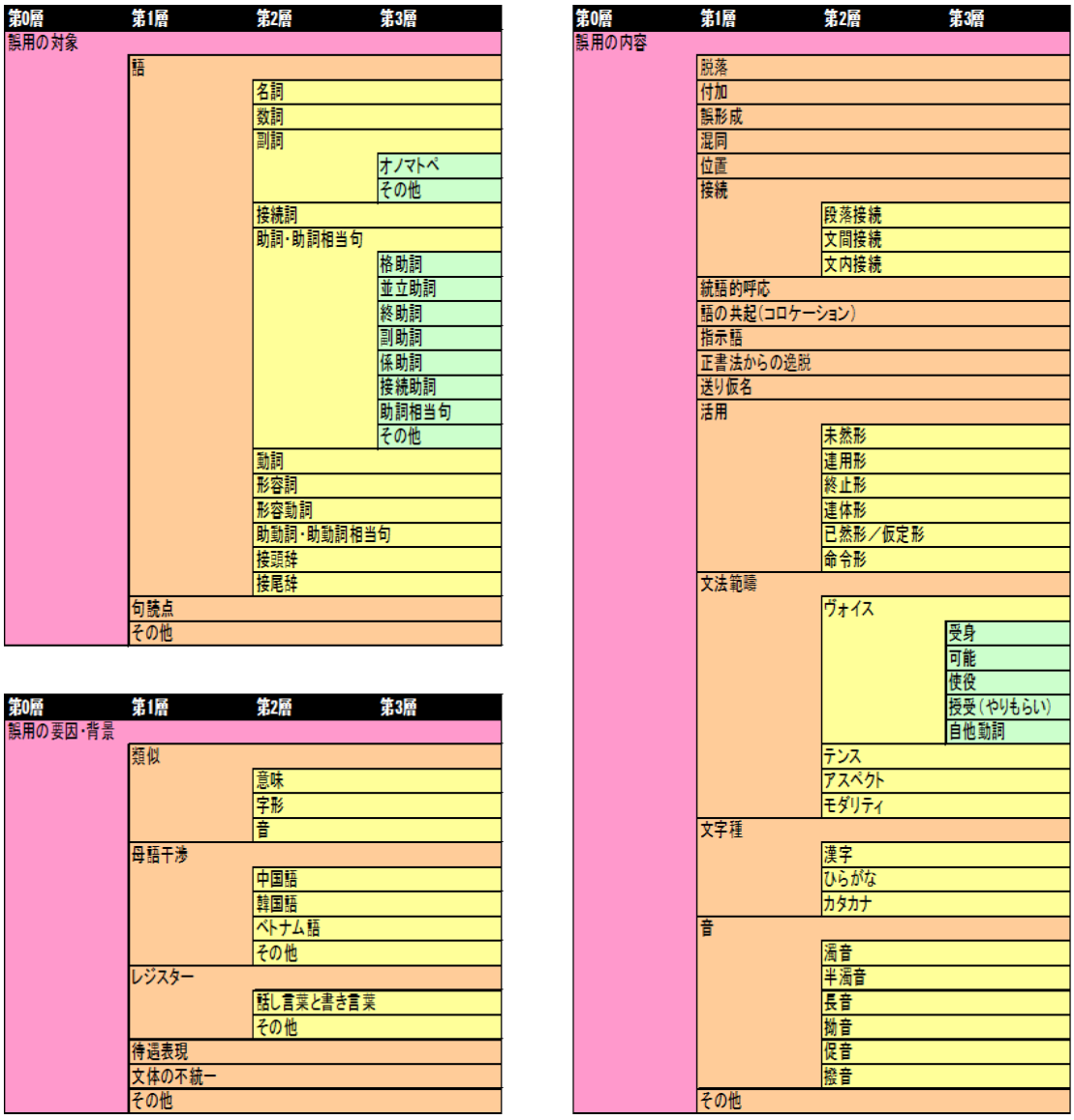

図１　策定した誤用タグの階層構造

当初は、それぞれがすべての作文を対象に共有できる策定案に従って、試験的にタグを付け、その後、再び妥当性を検討し、1) 誤用の対象、2) 誤用の内容、3) 誤用の要因・背景に分けて、誤用を多層的に観察し、タグ付けすることにした。

3.1 誤用の対象

「誤用の対象」は文中のどの部分が問題かを指摘する。図 1 の左上の図でその全体を示す。文全体の構造として文、句、節の単位を示す「句読点」の正誤を判定する。さらに、その下位の分類としては、品詞ごとに観察する。単語の組み合わせの例では、「表記をひらがなだけにした方がよくないと考える。(086_a)」(「にしない方がいい」) のように意味的かつ構文的に不適切なものが誤用対象として取り上げられる。

3.2 誤用の内容

図 1 の右側にある「誤用の内容」は対象となる部分の何が問題かを指摘する。図中の上部 5 項目「脱落・付加・語形成・混同・位置」は、対象となる箇所の誤用の種類を指摘する。この 5 項目に続く「接続」以後は、何が誤用なのかを指摘している。例えば、「話し合うことがだい好き__からです (P34_a)」は形容動詞語幹に続く助動詞「だ」の「脱落」の例として指摘される。「私はしょうらいにいろいろな事もしたいいんです。(p15_a)」では、助詞「に」、また「したいんです」はひらがな「い」が「付可」した例として指摘できる。

3.1 で見た「ひらがなだけにした方がよくない (086_a)」は、「しない (否定) 方が＋よい (否定形)」という定型的な表現を反対にした「混同」の例といえる。また、「何か外国語を習うか」と思って、「日本語を習うことにしました。(p34_a)」は、句および文の切れ続きが不適切で、「接続」「文間接続」にも該当している。

さらにまた、「したいいんです」は「文字」の書き間違い、あるいは「音」の誤認識の問題としても捉えられる。これらの現象は、その背景に学習者の文法誤りあるいは音声の語認識など要因として、図 1 左側下方にある「誤用の要因・背景」の項目が考えられる。

3.3 誤用の要因・背景

図 1 の「誤用の要因・背景」は、誤用となる原因および何故そうなるか、その背景を推測するものである。「類似」「母語干渉」「レジスター」「待遇表現」「文体の不統一」などが含まれる。

「日本の成功の経験を参考して、自分の国へもたらす国もあります。(159_a)」は中国語母語話者の例であり、「参考にする」の助詞「に」が脱落している。これは「に」を付加しないサ変名詞と考えるか、「名詞」と考えるかの判断によるが、中国語母語話者にとっては、同形の 2 字漢語があるため、その区別が困難になることから、母語干渉と考えられる。

「たくさん勉強もすることができます。石の上に三年ね。今日習いて、明日良い人にな

ります。」（P19_a）は話しことばと書きことばを混同した「レジスター」の例である。

４．誤用タグ付き作文と学習者属性データ XML の概要

作文と誤用タグデータは natane.xml に、学習者データは learners.xml に収録されている。natane.xml の主な要素と属性は以下の通りである：

```
<database>
  <document name="001_a.txt" learner="001">
    <!-- 元作文のファイル名 100_a.txt と学習者 ID (21) -->
    <text><!-- 作文の文字列（CDATA セクション） --></text>
    <segment start="10" end="16">
      <!-- text 内の誤用箇所の開始文字位置 start と終了文字位置 end
           下記 attribute は該当するもののみ記載 -->
      <attribute key="10-訂正例">
        <value>
          <!-- 誤用開始位置 start・終了位置 end に対する訂正例 -->
        </value>
      </attribute>
      <attribute key="20-対象">
        <!-- 誤用の対象に関する情報 -->
        <value><!-- 誤用タグは複数可能 --></value>
        <value><!-- 各層が::で区切られている --></value>
      </attribute>
      <attribute key="30-内容"><!-- 上記同様 --></attribute>
      <attribute key="40-要因・背景"> <!-- 上記同様 --></attribute>
      <attribute key="60-備考"> <!-- 上記同様 --></attribute>
      <content><!-- 参照用誤用文字列（CDATA セクション） --></content>
    </segment>
  </document>
</database>
```

natane.xml は各 document の learner 属性により、learners.xml の learner の id と結び付ける。下記のように、学習者識別子（001）に対し、母語、国籍、性別（一部）を記述している。learner.xml の要素と属性は以下の通りである：

```
<learners>
```

```
  <learner id="001">
    <gender><!-- 性別 --></gender>
    <nationality><!-- 国籍 --></nationality>
    <native_language><!-- 母語 --></native_language>
  </learner>
</learners>
```

3.1 で例にした「表記をひらがなだけにした方がよくないと考える。(086_a)」の下線部分を誤用タグで表すと以下の通りになる。

```
<segment id="17390" start="156" end="164" tag_name="error">
  <attribute key="30-内容">
    <value>語の共起(コロケーション)</value>
  </attribute>
  <attribute key="10-訂正例">
    <value>しない方がいい</value>
  </attribute>
  <attribute key="60-備考">
    <value>した方がよくない→しない方がいい</value>
  </attribute>
  <content><![CDATA[した方がよくない]]></content>
</segment>
```

本文の 156 文字目から 164 文字目までの誤用範囲「した方がよくない」に対し「しない方がいい」という訂正が与えられ、誤用の内容が「語の共起(コロケーション)」になっている。

５．Slate の利用

Slate は徳永・Kaplan によって言語コーパスにアノテーションを付与するために開発された自然言語処理技術のためのツールである（徳永健伸、Dain Kaplan、飯田龍（2010）。英語の他に多言語に付与することが可能で、日本語データを扱う「なたね」でも実施できる。このツールでは、人手で誤用対象（誤用の対象）を選んだ後は、その対象がどのような誤用であるか（誤用の内容）をあらかじめリストで提供された項目を選択し、誤用の要因も準備された候補から選択するというシステムである。そこでは、自由記述による恣意性が排除され、複数アノテータによるアノテーションでも表現の統一が保たれる。

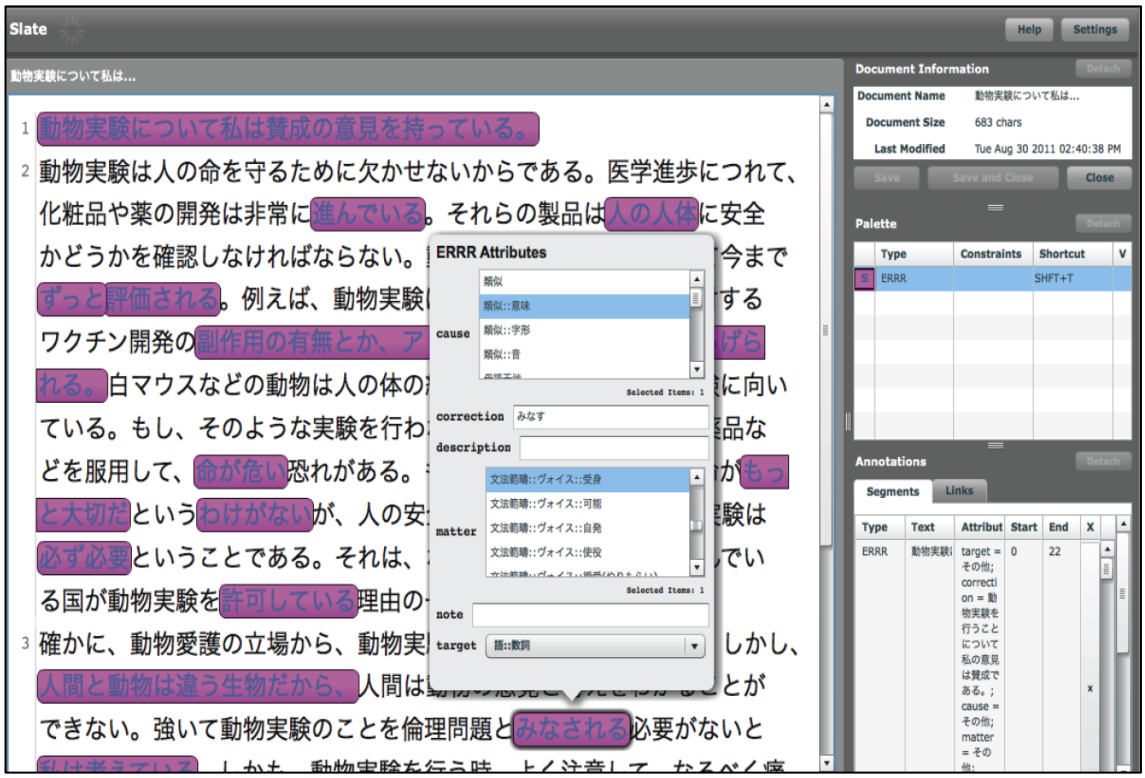

図 2　アノテーションツール Slate の入力画面（Dain Kaplan et al. 2011 の図から引用）

図 2 はある日本語学習者作文の一部である。アノテータは画面の第 1 行目の「動物実験について私は賛成の意見を持っている。」を「誤用の対象」として選択し、画面の右下の segments 記述欄に修正案を「私の意見は賛成である」と記述している。その次のプロセスとして、図の中央下部に移動し、「倫理問題とみなされる必要がないと私は考えている。」という箇所で「みなされる」を誤用対象として選択し、「誤用の内容」の箇所で「文法範疇::ヴォイス::受身」を選択している。

このように Slate というツールを利用することで、個人個人のアノテータの負担が軽減され、複数のアノテータで作業する場合は、見解の不一致も減少させることができた。

6．統計情報

本節では、「なたね」データに関する統計情報を示す。

6.1　母語別学習者数および作文数

表 1 母語別学習者数・作文数

母語	学習者数	作文数
中国語	115	152
マラーティー語	36	36
ベトナム語	13	27

韓国語	11	34
スペイン語	2	2
マレー語	1	8
スロベニア語	1	7
ハンガリー語	1	1
タイ語	1	1
母語未入力	11	17
計	192	285

注：学習者数と作文数が異なるのは、一人の学生が複数の作文を書いているためである。

6.2　誤用の対象

項目	項目数
語	
語::名詞	730
語::数詞	9
語::副詞	164
語::副詞::オノマトペ	4
語::副詞::その他	13
語::接続詞	95
語::助詞・助詞相当句	1058
語::助詞・助詞相当句::格助詞	201
語::助詞・助詞相当句::並立助詞	1
語::助詞・助詞相当句::終助詞	5
語::助詞・助詞相当句::副助詞	1
語::助詞・助詞相当句::係助詞	53
語::助詞・助詞相当句::接続助詞	26
語::助詞・助詞相当句::助詞相当句	28
語::助詞・助詞相当句::その他	11
語::動詞	751
語::形容詞	176
語::形容動詞	37
語::助動詞・助動詞相当句	223

語::接頭辞	0
語::接尾辞	11
句読点	209
その他	80
計	3886

6.3　誤用の内容

誤用の内容	項目数
脱落	161
付加	130
誤形成	234
混同	833
位置	38
接続	21
接続::段落接続	1
接続::文間接続	22
接続::文内接続	178
統語的呼応	229
語の共起(コロケーション)	199
指示語	49
正書法からの逸脱	7
送り仮名	23
活用	
活用::未然形	12
活用::連用形	88
活用::終止形	31
活用::連体形	14
活用::已然形／仮定形	3
活用::命令形	1
文法範疇	
文法範疇::ヴォイス	
文法範疇::ヴォイス::受身	76
文法範疇::ヴォイス::可能	47
文法範疇::ヴォイス::自発	11

文法範疇::ヴォイス::使役	23
文法範疇::ヴォイス::授受(やりもらい)	8
文法範疇::ヴォイス::自他動詞	42
文法範疇::ポラリティ	0
文法範疇::テンス	113
文法範疇::アスペクト	127
文法範疇::モダリティ	76
文字種	
文字種::漢字	97
文字種::ひらがな	108
文字種::カタカナ	69
音	
音::濁音	94
音::半濁音	3
音::長音	56
音::拗音	5
音::促音	29
音::撥音	4
その他	405
計	3667

6.4　誤用の要因・背景

項目	項目数
類似	
類似::意味	242
類似::字形	55
類似::音	133
母語干渉	46
母語干渉::中国語	4
母語干渉::韓国語	3
母語干渉::ベトナム語	2
母語干渉::その他の母語	3
レジスタ	

レジスタ::話し言葉と書き言葉	481
レジスタその他	2
待遇表現	8
文体の不統一	478
その他	13
計	1470

参考文献

曹紅荃・黒田史彦・八木豊・鈴木泰山・仁科喜久子（2010）「学習者作文支援システムのための誤用データベース作成－動詞の誤用分析を中心に－」『世界日語教育大会論文集』, 1571-1-1571-9, 2010 年 8 月

曹 紅荃, 仁科 喜久子 (2011)「誤用データベースにおける誤用種別の策定（Establishment of Error Classification Framework for Error Database)」 日本語教育方法研究会誌、Vol.18, No.1, pp. 38-39, 2011 年 3 月

曹紅荃・黒田史彦・八木豊・仁科喜久子（2011）「学習者作文コーパスのための誤用種別の整備と分析」《跨文化交际中的日语教育研究②》『異文化コミュニケーションのための日本語教育②』、高等教育出版社, 520-521, 2011 年 8 月

曹 紅荃、黒田 史彦、八木 豊、仁科 喜久子(2012)「学習者コーパス『なたね』の構築と応用の可能性（Construction of Learner corpus “NATANE” and possible application)」『第 5 回「日本語教育とコンピュータ」国際会議 Castel/J 予稿集（Proceedings of the 5th International Conference on CASTEL-J (Computer Assisted Systems For Teaching & Learning Japanese)』Poster 5: pp. 1-4, 2012CASTEL-J, 2012 年 8 月

曹紅荃、仁科喜久子（2012）「学習者作文コーパスのための誤用種別標準化に向けての検証と確定」『2012 年日本語教育国際研究大会予稿集第 1 分冊』P58, 社団法人日本語教育学会, 2012 年 8 月

徳永 健伸、Dain Kaplan、飯田 龍（2010）「Slate - A multi-purpose annotation tool」『情報処理学会自然言語処理研究会報告』 情報処理学会, NL-199, 19

仁科 喜久子、八木 豊、阿辺川 武、ホドシチェク ボル（2017）『習ったはずなのに使えない文法』くろしお出版, pp.211-232

八木 豊、鈴木 泰山、仁科 喜久子（2011）「BCCWJ と誤用コーパスを利用した日本語作文支援に関する一考察」『特定領域研究日本語コーパス平成 22 年度公開ワークショップ（研究成果報告会）予稿集』 文部科学省科学研究費特定領域研究「日本語コーパス」総括班, pp.119-124

八木 豊、ホドシチェク・ボル、仁科 喜久子(2012)「BCCWJ と学習者作文コーパスを利用した日本語作文支援-表記と共起に関する誤用添削プロトタイプ構築」第 1 回コーパス日本語学ワークショップ, 国立国語研究所

八木 豊、Bor HODOŠČEK、阿辺川 武、仁科 喜久子（2013）「学習者が犯す誤用の要因・背景からみる日本語作文支援」第 3 回コーパス日本語学ワークショップ, 国立国語研究所

Dain Kaplan, Ryu Iida, Kikuko Nishina, Takenobu Tokunaga (2011) Slate-A tool for Creating and Maintaining Annotated Corpora, JLCL2011-Band26(2), pp.89-101

科学研究費補助金

共同研究・競争的資金等の研究課題

1）科学研究費補助金　基盤研究（B）「大規模コーパスを利用した日本語学習支援システム『ひのき』構築と評価研究機関」研究代表者　仁科 喜久子（東京工業大学留学生センター）研究期間（2009～2011）

2）科学研究費補助金　挑戦的萌芽研究「日本語学習者誤用コーパスを利用した作文システムの開発」研究代表者　仁科 喜久子（東京工業大学留学生センター）研究期間（2010～2012）

3）科学研究費補助金　基盤研究（C）「日本語作文支援システムで考慮すべき学習者属性情報と提示項目の分析研究」研究代表者　阿辺川 武（国立情報学研究所）研究期間（2012～2014）

開発者および担当者名　ひのきプロジェクト（2021 年 10 月現在）

阿辺川武（国立情報研究所）

八木豊（株式会社　ピコラボ）

ボル・ホドシチェック（大阪大学）

仁科喜久子（東京工業大学　名誉教授）（連絡担当者）

担当者の Email アドレス　knishina@m06.itscom.net