Vol.2021-MUS-131 No.31
Vol.2021-SLP-137 No.31
2021/6/18

ITA コーパス：パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価

小口 純矢[1,a)]　金井 郁也[1,b)]　小田 恭央[2,c)]　齊藤 剛史[3,d)]　森勢 将雅[1,e)]

概要：音素バランスを考慮したパブリックドメインの日本語テキストコーパスである ITA コーパスを構築した．無償で利用できる従来のコーパスは「継承」を条件に公開されていることが多い．そのため，二次著作物を同じライセンスのもとで公開しなければならず，商用利用や公開範囲を制限したい場合に不便である．また，限られた文数のなかで音素バランスを考慮した結果，読みにくい文章が多くなり，話者の負担となってしまう．ITA コーパスは著作権の消滅した文献あるいはオリジナルの文章・単語から文セットを構築することでパブリックドメインで公開される．100 文と 324 文のサブセットからなり，それぞれが日本語における可能なモノフォン・ダイフォンを豊富に含んでいる．また，音素バランスだけでなく読みやすさも考慮されており，用途・分野を問わず幅広い応用が可能である．実際の応用事例として，感情音声コーパスおよび読唇データベースを構築した例も紹介する．

キーワード：コーパス，データベース, 音素バランス, マルチモーダル，音声合成

1. はじめに

統計的音声合成の研究において利用されるコーパスの設計は，そのコーパスを用いて訓練された合成器の性能に大きく影響を与える．一般に，統計的生成モデルは学習データに含まれない音韻の合成は困難である．そのため，日本語において出現する音韻をなるべく多様に含むように設計された音素バランス*1文の存在は欠かせないものとなっている．また，既存の音声コーパスをカバー・改変し自らの声でパラレルコーパスを収録し，新たなデータベースとして公開する試みも盛んに行われており，豊富な音声資源を利用して音声研究はますます活性化していくと思われる．

一方で，既存の音素バランス文セットは，誰もが自由に利用可能というわけではない．ATR 音素バランス文 [1] はエントロピーに基づく文選択など綿密に設計された信頼性の高いコーパスであるが，有償で提供されており，新たに収録した読み上げ音声を公開することができない．また，声優統計コーパス [2] や JSUT [3] は無償で利用できるものの，「継承」がライセンス条項に含まれており二次著作物を同様のライセンスのもとで頒布しなければならない．そのためコーパスの利用者は，作成した音声合成システム・音声コーパスを同じライセンスのもとで公開しなければならず，商用利用や利用範囲を制限することができない．また，既存の音素バランス文は限られた文数で音素バランスのみに基づいた文選択を行っている．そのため，読み慣れない単語や長い文が多く含まれる傾向にあり，アマチュアの話者にとって正しく発話することが難しい場合がある．

これらの問題を解決と，音声分野のみならず関連研究のさらなる活性化を目的に「分野横断的研究を加速させる」コーパスとして Inter-field Task Accelerating; ITA コーパスを構築した．本コーパスは以下の特徴がある．

パブリックドメイン: 著作権切れの作品をもとに作成されているため，完全に著作権フリーで公開される．これにより，利用者は商用・非商用問わず二次著作物の頒布を自由に行うことができる．

音素バランス: 日本語において可能なモノフォン・ダイフォンを全て，トライフォンをできるだけ豊富に含むように設計されている．

読みやすさ: 読みやすさを考慮して文の選定・改変を

[1] 明治大学, Meiji University, Nakano, Tokyo, 164–8525, Japan
[2] SSS 合同会社, SSS LLC., Sendai, Miyagi, 983–0831, Japan
[3] 九州工業大学, Kyushu Institute of Technology, Izuka, Fukuoka 820–8502, Japan
[a)] cs202027@meiji.ac.jp
[b)] cs202003@meiji.ac.jp
[c)] oda@zunko.jp
[d)] saitoh@ai.kyutech.ac.jp
[e)] mmorise@meiji.ac.jp

*1 ここでいう音素バランスとは，「音素・音素連鎖がなるべく多種類・等確率で含まれる」ことを指し，「この世界の日本語文章全てを母集団とした音素の出現確率の分布に近い」という意味ではないことに注意されたい．

行っており，収録時の話者の負担を低減するように設計されている．

以降では，ITA コーパスの設計と分析結果，およびその応用事例である感情音声コーパスと読唇用マルチモーダルデータベースについて報告する．

2. コーパスの設計

2.1 構成

ITA コーパスは 424 文の日本語文章からなる音素バランス文セットである．モノフォン・ダイフォンを考慮した 100 文からなる Emotion セットと，モノフォン・ダイフォンに加えてトライフォンを豊富に含むように作られた 324 文からなる Recitation セットの 2 つから構成される．これにより，複数話者によるパラレルデータや感情音声の収録には文数の少ない Emotion セットを利用し，比較的収録時間に余裕がある場合には Recitation セットも合わせて用いるなど，サブセット毎に独立して音素バランスが担保されているため，目的に合わせて使い分けることができる．

2.2 コーパス文の選定

青空文庫 [4] に収蔵されている著作権切れの文学作品，およびパブリック・ドメインの日英対訳コーパスである田中コーパス [5] から，担保すべき音素連鎖を含む文章をランダムに選択した．その後，読みやすさと公序良俗に照らして以下の項目に該当する文について削除・改変を行った．

- 馴染みのない外来語・地名など読みにくい単語を含む
- 商品・サービス・社名を含む
- 差別的・性的・暴力的な表現を含む

以下は，読みやすさを考慮して文の改変を行った例である．

修正前：「六月二十日に民衆をテュルリー宮殿に走らした肉屋のレジャンドルも、王を廃するということは夢にも考えなかった。」

修正後：「レジャンドルは民衆をテュルリー宮殿に招いた。」

修正前の文は，外国の人名・地名が含まれているだけでなく，文が長いため言い淀みやすい．一方，改変後の文は必要な音素連鎖を含みつつ文が短くなっており読みやすくなっている．

3. コーパスの分析

ITA コーパスの音素バランスについて，既存の音素バランス文と比較しながら分析を行った．分析対象とするのは以下の 5 セットである．

- **ATR503**: ATR 音素バランス文（503 文） [1].
- **声優統計**: 声優統計コーパス（100 文） [2].
- **ITA-emo**: ITA コーパス Emotion セット（100 文）.
- **ITA-recit**: ITA コーパス Recitation セット（324 文）.
- **ITA-full**: ITA コーパス全体（424 文）.

Table 1: 各コーパスに含まれるトライフォンの種類数．（モノフォン：39 種類，ダイフォン：459 種類は共通）．

Corpus	Triphone [種類]
ATR503	2976
声優統計	2096
ITA-emo	1686
ITA-recit	2839
ITA-full	3155

3.1 音素連鎖の種類

まず，各コーパスに含まれるトライフォンの種類数を計算した．**Table 1** はその結果である．ITA-full が最も多くの種類のトライフォンを含んでいるという結果が得られた．これは ATR503 がエントロピーを基準に文を選択しており，ITA コーパスは種類数を基準に選択していることが原因と考えられる．

続いて，ITA-emo・ITA-recit について子音-母音（Consonant–Vowel; CV）連鎖の出現数を計算した．それぞれの計算結果を **Table 2**・**Table 3** に示す．どの CV 連鎖も全体として 2 つ以上含まれていることが分かる．

3.2 エントロピー

次に，各音素連鎖の出現頻度が一様であるかを示す指標として，以下によって定義される拡張エントロピーを計算した [6].

$$S = \sum_{m=1}^{M} w_m S_m, \quad S_m = -\sum_{n=1}^{N_m} p_{mn} \log_2 p_{mn} \tag{1}$$

ここで，M は音素連鎖の数であり，モノフォンは 1，ダイフォンは 2，トライフォンは 3 を表す．S_m と N_m はそれぞれ m 番目の音素連鎖とその総数である．p_{mn} は音素連鎖 n がコーパス内で現れる確率である．w_m は各音素連鎖に対する重みであり，本研究ではいずれも 1.0 に設定した．拡張エントロピーは各音素連鎖が同じ割合で含まれるほど大きな値をとる性質を持つため，コーパスの音素バランスを定量的に表すことができる．計算結果を **Table 4** に示す．全体の傾向として文数が増えるほどエントロピーも増大することが読み取れる．一方で，声優統計の方が同じ文数である ITA-emo よりも大きな値を取っている．これは読みやすさを考慮して文を修正したことに起因すると考えられる．

4. 応用例

本章では，ITA コーパスの利用事例として，感情音声コーパスおよび読唇用マルチモーダルデータベースを構築した例を紹介する．

Table 2: Emotion セットにおける CV 連鎖の出現頻度
（空欄 – は日本語において不可能な連鎖）

		Vowel				
		a	i	u	e	o
Consonant	b	13	4	14	9	4
	by	1	—	1	1	4
	ch	2	31	1	1	2
	d	30	4	1	44	30
	dy	1	—	4	—	1
	f	2	2	11	1	3
	g	62	6	4	12	10
	gy	1	—	2	1	4
	h	17	16	—	4	12
	hy	1	—	1	1	5
	j	3	25	2	1	15
	k	96	40	69	27	51
	ky	1	—	3	1	8
	m	42	20	10	12	36
	my	1	—	2	2	4
	n	62	62	2	11	106
	ny	2	—	5	1	1
	p	6	2	4	3	5
	py	1	—	6	1	2
	r	42	29	71	39	7
	ry	1	—	2	1	6
	s	19	1	64	17	18
	sh	6	84	8	1	10
	t	102	4	1	68	98
	ts	2	1	32	1	1
	ty	1	—	3	—	1
	v	2	1	1	1	1
	w	99	1	—	1	1
	y	7	—	10	1	24
	z	6	1	12	5	2

Table 3: Recitation セットにおける CV 連鎖の出現頻度
（空欄 – は日本語において不可能な連鎖）

		Vowel				
		a	i	u	e	o
Consonant	b	32	10	27	16	14
	by	2	—	2	1	7
	ch	16	55	23	3	22
	d	78	5	1	134	48
	dy	1	—	4	—	1
	f	12	9	67	9	3
	g	156	14	25	26	24
	gy	6	—	8	1	17
	h	69	48	—	18	35
	hy	12	—	4	1	15
	j	6	44	13	2	25
	k	214	104	179	51	123
	ky	13	—	23	1	38
	m	116	56	27	42	87
	my	1	—	2	1	6
	n	154	184	12	27	268
	ny	10	—	12	1	3
	p	27	13	22	10	26
	py	2	—	3	1	8
	r	111	75	186	87	31
	ry	1	—	9	1	23
	s	84	1	156	61	51
	sh	26	229	30	5	52
	t	271	14	3	147	229
	ts	3	2	113	2	1
	ty	1	—	5	—	1
	v	3	1	1	4	3
	w	227	4	—	4	7
	y	40	—	28	1	50
	z	14	1	32	12	9

4.1 感情音声コーパス

100 文セットを感情音声，324 文セットを通常の読み上げ音声としてそれぞれ利用し，感情音声コーパスを構築した．既存の感情音声コーパスには，怒り・悲しみ・喜びといった典型的な感情音声が収録されていることが多いが，本音声コーパスはキャラクターボイスとしての用途を想定し，あまあま（甘えた声）・セクシー（大人っぽい声）・ツンツン（いわゆるツンデレ声）の 3 感情を収録した．加えて，それぞれに対してプロのアノテータによって付与された音素境界の時間アライメントに基づくフルコンテキストラベルを付属する予定である．

4.2 読唇マルチモーダルデータベース

音声に加えて動画像を補助特徴量として用いることで精度向上を目指す試みが古くからなされている．特に深層学習によるアプローチは大量のデータが必要であるため，異なるドメインのメディアを収録したマルチモーダルデータベースには需要が見込まれる．例えば，口唇の動画像から音声認識・合成を行うタスクは，聴覚・発話障害者のコミュニケーションを支援する上で重要な技術であることから，技術研究のみならずデータベースの整備も盛んである [7]．

Table 4: 各コーパスの拡張エントロピー．

Corpus	Extended entropy [bit]
ATR503	22.49
声優統計	22.27
ITA-emo	21.85
ITA-recit	22.34
ITA-full	22.38

こうした背景から，我々は上の感情音声コーパスに収録中の口唇の動きを記録した動画像データを付録させて読唇用データベースの公開を予定している．

4.3 さらなる応用可能性

ITA コーパスは，その読み上げ音声を音声認識・合成システムの学習データとして利用したり，新たなデータベースの構築に活用するなど，典型的な用途だけでも幅広い応用を考えることができる．

また，ITA コーパスに含まれる音素環境の豊富さから，その読みやすさを前後の音韻の関係から多角的に分析することができ，speech error（言い間違い）や speakability（発話しやすさ）に関する研究への応用が期待できる．

5. おわりに

本研究は, パブリックドメインの音素バランス文セットである ITA コーパスを構築した．3 音素連鎖だけでなく読みやすさを考慮しており，幅広い用途で利用可能である．音素バランス・エントロピーに基づく評価において ITA コーパスの有効性が示された．また，感情音声コーパスや読唇マルチモーダルデータベースといった応用事例を紹介した．これらは研究用データベースとして公開を予定している．今後は読みやすさについての評価や音声合成に利用した場合の性能比較を行っていくことを考えている．

謝辞 本研究は，科研費 JP21H04900 の支援を受けた．また，文章の選定にご協力いただいた細田計さんに感謝の意を示す．

参考文献

[1] 磯 健一, 渡辺 隆夫, and 桑原 尚夫, “音声データベース用文セットの設計,” 1988, pp. 89–90.
[2] y_benjo and MagnesiumRibbon, “声優統計コーパス,” http://voice-statistics.github.io.
[3] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” *abs/1711.00354*, Nov. 2017.
[4] “青空文庫,” https://www.aozora.gr.jp/.
[5] 田中康仁, “田中コーパス（パブリック・ドメイン版）.”
[6] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, “Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical parametric speech synthesis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1107–1116, 2017.
[7] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” in *Proc. Interspeech*, 2018, pp. 3244–3248.