2-2-2　対象言語の諸相と多言語化への支援体制

2-2-2　Aspects of the Target Languages and the Supporting Framework for Multilingualization

加藤 宏明
KATO Hiroaki

本稿執筆時点（2022 年 11 月）で NICT ASTREC における自動音声認識（ASR）・テキスト音声合成（TTS）の対象は 21 言語であった。本稿では、研究開発においてそれら対象言語の全体像を把握するために共有していた表データを掲載するとともに、多言語化を効率的に進めるための研究室の体制について報告する。

At the time of this writing (November, 2022), the number of target languages for automatic speech recognition (ASR) and text-to-speech synthesis (TTS) at ASTREC, NICT was 21. This article describes the tabular data that we shared among the lab members to get an overall picture of these target languages, as well as the supporting framework at our lab which efficiently promotes multilingualization.

1　まえがき

音声翻訳は、自動音声認識（ASR）を入力、テキスト音声合成（TTS）を出力として異なる言語を話す人々相互のコミュニケーションを可能とする技術である。多言語化によりこのコミュニケーションの輪は劇的に広がる。これまでに当研究室が手がけた 21 言語は表 1 のとおりだが、『話者数』を見ると、新しい言語が加わるごとに五百万人から十四億人が輪に入ってきた勘定になる。一方で、同じ表は世界の言語がおそろしく多様であることも示し、多言語化が簡単な仕事ではないことを教える。幸いなことに、近年の人工知能に基づく音声の処理方法は、原理的には言語に依存しない手法が主流である。しかしながら、これに携わる者が対象言語の知識を一切持たなくて済むかと問われれば、現実的に利用可能なデータの質と量を考えれば、まったくそうではない。この、対象言語に関わる部分への当研究室の取組を、非母語話者の開発者が知っておくべき基礎知識と母語話者でしか担えない役割の両面から述べる。

2　対象言語の諸相

新しい対象言語の追加が計画されると表 1 に行が追加される。多言語音声翻訳に関わるメンバー間での情報共有のためである。項目（列）はメンバーの要求に応じて徐々に増えた。以下に各項目の意図を概説する。

『言語名』と『話者数』は最も基本的な情報であり、新たに輪に加わる人口の規模を示す。続く『系統、類型、基本語順』は『正書法の文字』とともに言語間の類似性を示す手がかりであり、これらが一致する言語同士では処理系の共有や軽微な改変でのツールの流用が期待できる。『音素数、声調』は単語の識別に必要な音の要素数を示し、G2P（grapheme-to-phoneme convertor）などの基本的なツールの設計に役立つ。『標準語、方言分布』は主要な方言とその地域分布を示し、音声コーパスを収集すべき地域や範囲の検討に資する。これに続く正書法に関する諸情報はテキスト処理系の選択や設計に資するとともに、テキストデータの自動品質チェックでも使われる。『ラテン文字公式表記法』はラテン文字以外を正書法で使う言語に与えられる情報で、これだけで簡易版の G2P として使える場合もあるので含めている。

次項目以降は、ASR もしくは TTS の開発側から要求があったものである。『語彙表現の性差』は話者の性別への表現依存性の有無で、TTS 用の文生成における語彙選択で考慮しなければならない。『数字の読みの規則性』と『TTS との親和性』も TTS 用の情報である。特に、『数字の読みの規則性』は TTS にとっては古典的な課題であり、これが低い場合はそれなりの覚悟が求められる。たとえば日本語や韓国語は、漢語・固有語の 2 種類の読み方が混在し、後続する助数詞によっても

表 1　対象 21 言語の諸相

(*n は脚注番号を示す)

項目 / 言語名 *1	母語話者数 *2	言語系統 *3	形態の類型	基本語順 *4	音素数 *5	声調	標準語 *6	方言分布 *7	正書法の字種 *8	正書法の文字と書字方向 *9
日本語 **にほんご** **ja, jpn, 日**	125 M (125 M, L1+L2)	日琉諸語	膠着語	SOV AN	5 母音×長短 14 子音	なし	東京方言	東日本 (60%) 西日本 (39%) 他	表意文字と表音文字が混在	漢字+ 2 種の固有文字 (仮名) 左→右, 上→下 (右先頭)
英語 **English language** **en, eng, 英**	373 M (1452 M, L1+L2)	印欧語族 - ゲルマン語派	屈折語	SVO AN	11 母音 24 子音	なし	国ごとに異なる	北米 (74%) 欧州 (17%) 大洋州 (5%) 他	表音文字	ラテン文字 左→右
中国語 [3.a] **汉语** **zh, zho/chi, 中**	930 M (1118 M, L1+L2)	シナ・チベット語族 - シナ語派	孤立語	SVO AN	6 母音 21 子音	4 声調 +軽声	北京方言	華北東北 (30%) 西南 (30%) 西北 (20%) 江淮 (10%) 他	表意文字	漢字 (簡体字, 繁体字) 左→右, 上→下 (右先頭)
韓国語 **한국어** **ko, kor, 韓**	82 M (韓国 51 M)	朝鮮諸語	膠着語	SOV AN	9 母音 19 子音	なし	ソウル方言	中部 (62%) 東南 (22%) 西南 (14%) 他	表音文字	固有文字 (ハングル) 左→右, 上→下 (右先頭)
タイ語 **ภาษาไทย** **th, tha, 泰**	21 M (61 M, L1+L2)	タイ・カダイ語族	孤立語	SVO NA	9 母音×長短 21 子音	5 声調	中部 (バンコク) 方言	中・東部 (37%) 北東部 (34%) 南部 (14%) 北部 (10%) 他	表音文字	固有文字 Brahmic (タイ文字) 左→右
ベトナム語 **tiếng Việt** **vi, vie, 越**	85 M (85 M, L1+L2)	オーストロアジア語族 - モン・クメール語派	孤立語	SVO NA	11 母音 19 子音	6 声調	北部 (ハノイ) 方言	北部 (40%) 中部 (8%) 南部 (52%)	表音文字	ラテン文字+付加記号 (5 声調記号) 左→右
インドネシア語 **bahasa Indonesia** **id, ind, 尼 (印尼)**	44 M (199 M, L1+L2)	オーストロネシア語族 - マレー・ポリネシア語派	膠着語	SVO NA	6 母音 18 子音	なし	あり	インドネシア語 (20%) ジャワ語 (32%) スンダ語 (15%) 他	表音文字	ラテン文字 左→右
ミャンマー語 **မြန်မာဘာသာ** **my, mya/bur, 緬**	33 M (43 M, L1+L2)	シナ・チベット語族 - チベット・ビルマ語派	膠着語	SOV AN	8 母音 23 +11 子音	3 声調	中部 (ヤンゴン・マンダレー) 方言	上流部 (41%) 下流部 (39%) 東部 (11%) 西部 (6%) 他	表音文字	固有文字 Brahmic (ミャンマー文字) 左→右
フランス語 **français** **fr, fra/fre, 仏 (法)**	80 M (274 M, L1+L2)	印欧語族 - イタリック語派	屈折語	SVO NA	11 母音 (+3 鼻母音) 17 子音	なし	あり (放送語は国ごとに異なる)	仏北部 (56%) 仏南部 (22%) 北米 (12%) ベルギー (5%) 他	表音文字	ラテン文字+付加記号 (4 上部記号 , ç) 左→右
スペイン語 **idioma español** **es, spa, 西**	475 M (548 M, L1+L2)	印欧語族 - イタリック語派	屈折語	SVO NA	5 母音 16 子音	なし	国ごとに異なる	北米 (34%) 中南米カリブ海 (57%) 欧州 (9%) 他	表音文字	ラテン文字+付加記号 (2 上部記号) 左→右
ポルトガル語 **língua portuguesa** **pt, por, 葡**	232 M (258 M, L1+L2)	印欧語族 - イタリック語派	屈折語	SVO NA	8 母音 (+5 鼻母音) 19 子音	なし	国ごとに異なる	ブラジル (86%) ポルトガル (5%) アフリカ (9%) 他	表音文字	ラテン文字+付加記号 (4 上部記号 , ç) 左→右
フィリピン語 **wikang Filipino** **-, fil, 比 (菲)**	28 M (82 M, L1+L2)	オーストロネシア語族 - マレー・ポリネシア語派	膠着語	VSO NA	5 母音 20 子音	なし	マニラ首都圏方言	タガログ語 (29%) セブ語 (23%) 他	表音文字	ラテン文字 左→右

正書法の文字数 *10	正書法における語境界 *11	数字表記 *12	正書法補足（句読法など）*13	ラテン文字公式表記法	語彙表現の性差 *14	数字の読みの規則性 *15	TTS との親和性 *16	諾否疑問文の平叙文との違い *17
漢字：2,998 字 平仮名：48 字 片仮名：48 字	なし	アラビア数字 +漢数字 1234567890 一二三四五 …	文末に句点（. /。），文中に適宜読点（, /、）	ローマ字（訓令式 / 修正ヘボン式）	*	低	低	接尾語付加 （「か」） *
26 字×大小	スペース	アラビア数字 1234567890	文頭と固有名詞は大文字，文末記号は “./?/!”. ハイフン使用可	-	*	中	中	語順 / 接頭語付加
漢字：8,105 字	なし	アラビア数字 +漢数字 1234567890 一二三四五 …	文末記号は “。/？/！”. 読点は “, ”. 声調は正書法に現れない.	漢語拼音方案	なし	高	低	接尾語付加 （「吗」）
字母：24 種	スペース	アラビア数字 1234567890	文末記号は “./?/!”. 読点は “,”.	文化体育観光部 2000 年式ローマ字	なし	低	中	接尾語句変化 *
42 字	なし	アラビア数字 +タイ数字 1234567890 ๑๒๓๔๕๖๗๘๙๐	文末記号なし．声調は正書法で明示.	RTGS（タイ王立学士院式）	あり	高	中	接尾語句付加 （“ไหม” など）
29 字×大小 （7 独自文字：Ă, Â, Đ, Ê, Ô, Ơ, Ư）	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!”．声調は正書法で明示	-	**	高	高	機能語句（yes と no）付加（反復疑問が基本）
26 字×大小	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!” ハイフン使用可	-	なし	高	高	イントネーションのみ（文字では疑問符のみ）が多い
33 字	なし	ミャンマー数字 +アラビア数字 ၁၂၃၄၅၆၇၈၉၀ 1234567890	文末記号は “ ။ ”, 読点は “ ၊ ”．声調は正書法で明示	MLC（ミャンマー言語委員会）式	あり	中	中	接尾語付加 （“ လား ”）
28 字×大小 （2 独自文字：Æ, Œ）	スペース	アラビア数字 1234567890 小数点は “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!” ハイフン使用可	-	あり	高（性）	高	語順 / 機能語句付加．イントネーションのみのスタイルも広く流通
27 字×大小 （1 独自文字：Ñ）	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ , ”	文頭と固有名詞は大文字，文末にピリオド．疑問文，強調文は “¿ .. ?”, “¡ .. !” で囲む. ハイフン使用可	-	あり	高（性複）	高	イントネーションのみ（文字では疑問符のみ）
26 字×大小	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!”. 公式正書法はブラジル式 ハイフン使用可	-	あり	高（性複）	高	イントネーションのみ（文字では疑問符のみ）
28 字×大小 （2 独自文字：Ñ, Ng）	スペース	アラビア数字 1234567890	文頭と固有名詞は大文字，文末記号は “./?/!” ハイフン使用可	-	なし	低	高〜中	機能語付加（“ba” か “bang”） *

表 1　対象 21 言語の諸相

項目 / 言語名 *1	母語話者数 *2	言語系統 *3	形態の類型	基本語順 *4	音素数 *5	声調	標準語 *6	方言分布 *7	正書法の字種 *8	正書法の文字と書字方向 *9
クメール語 **ភាសាខ្មែរ** **km, khm**	16 M (16 M, L1+L2)	オーストロアジア語族 - モン・クメール語派	孤立語	SVO NA	9 短母音 +10 長母音 21 子音	なし	北西部 / プノンペン方言	首都圏 (28%) 北西部 (32%) 南部 (12%) 東部 (28%)	表音文字	固有文字 Brahmic (クメール文字) 左→右
ネパール語 **नेपाली भाषा** **ne, nep, (尼)**	16 M (25 M, L1+L2)	印欧語族 - インド語派	膠着語	SOV AN	5 短母音 +3 長母音 29 子音	なし	カトマンズ首都圏方言	ネパール語 (45%) マイティリー語 (12%) 他	表音文字	デーバナーガリ文字 Brahmic 左→右
モンゴル語 **Монгол хэл** **mn, mon**	5.2 M (モンゴル国 3 M)	アルタイ諸語 - モンゴル語族	膠着語	SOV AN	7 母音×長短 19 子音	なし	ハルハ方言	中央 (80%) 周辺 (20%)	表音文字	キリル文字 左→右 固有文字 (モンゴル文字) 上→下 (左先頭)
ロシア語 **русский язык** **ru, rus, 露 (俄)**	154 M (258 M, L1+L2)	印欧語族 - スラブ語派	屈折語	SVO AN	10 母音 (5 硬母音 +5 軟母音) 32 子音	なし	モスクワ方言	北部 (17%) 中部 (53%) 南部 (30%)	表音文字	キリル文字 左→右
アラビア語 **اللغة العربية** **ar, ara, 阿 / 亜**	口語 313 M 文語 274 M	アフロ・アジア語族 - セム語派	屈折語	VSO NA	3 母音×長短 28 子音	なし	現代標準アラビア語 (文語)	北部 (20%) 西部 (24%) 中央 (37%) 南部 (19%)	表音文字	アラビア文字 右→左
ドイツ語 **Deutsch** **de, deu/ger, 独 (德)**	95 M (180 M, L1+L2)	印欧語族 - ゲルマン語派	屈折語	SVO AN	14 母音 22 子音	なし	ハノーファーの方言 (諸説あり)	北部 (34%) 中部 (22%) 南部 (36%) 他	表音文字	ラテン文字+付加記号 (1 上部記号) 左→右
イタリア語 **lingua italiana** **it, ita, 伊 (意)**	65 M (68 M, L1+L2)	印欧語族 - イタリック語派	屈折語	SVO NA	7 母音 21 子音	なし	トスカーナ方言	北部 (42%) 中部 (18%) 南部 (30%) 他	表音文字	ラテン文字+付加記号 (2 上部記号) 左→右
ヒンディー語 **हिंदी** **hi, hin, (印)**	344 M (602 M, L1+L2)	印欧語族 - インド語派	膠着語	SOV AN	3 短母音 +8 長母音 34 子音	なし	デリー首都圏方言	ヒンディー語 (26%) ベンガル語 (18%) マラーティ語 (6%) テルグ語 (6%) 他	表音文字	デーバナーガリ文字+付加記号 (下点) Brahmic 左→右
ウクライナ語 **українська мова** **uk, ukr, 宇 (烏)**	40 M (45 M, L1+L2)	印欧語族 - スラブ語派	屈折語	SVO AN	10 母音 (5 硬母音 +5 軟母音) 32 子音	なし	キーウ首都圏方言	北部 (10%) 中央部 (17%) 東南部 (50%) 西部 (23%)	表音文字	キリル文字 左→右

*1　上から、日本語言語名、各言語自称、ISO の言語コード (ISO639-1、639-2 T/B)、日本 (中国) における言語の漢字略称を示す。「フィリピン語」はフィリピン共和国政府が国語として定める言語の呼称であり、言語学的には「タガログ語」に重なる。モンゴル語の固有文字は縦書きのみである。

*2　M: 百万人。L1 + L2：第 1 言語話者と第 2 言語話者の合計数。クメール語、ネパール語、モンゴル語、アラビア語口語、ドイツ語、ウクライナ語は Wikipedia 英語版か census、その他は Ethnologue (2022 年版) によった。アラビア語はアラビア語圏での共通語である現代標準アラビア語が開発対象である。基本的に文語のため、L1 話者はいない。

*3　言語系統、形態の類型は諸説ある。

*4　文要素や品詞の定義・機能は言語によって異なるが、基本的に右を想定した。S: 主語、O: 目的語、V: 動詞、A: 形容詞 (修飾側)、N: 名詞 (被修飾側)。あくまで基本的な語順であり、言語によってはこれに従わない文、表現もよく使われる。

*5　標準語を対象とした。日本語、英語、韓国語、タイ語、フランス語、ポルトガル語、ヒンディー語は IPA Handbook [1] を、その他は Wikipedia などを参考にした。母音・子音いずれも数え方には諸説ある。ここでは、二重母音、介音や介音との合成音素、異音は原則として含まない。母音の長短は、独立した 2 つの単母音とみなされる場合もある。二重母音を含めないのは、単母音の組み合わせで表現可能なものが多く、また複合母音 (母音連鎖) との区別が困難な場合があるためである。破擦音は、IPA で定義されていないためか、IPA Handbook の一部の言語では音素リストにないが、母語話者の直感を尊重してここでは含めた。ロシア語、ウクライナ語の硬音・軟音の対立は、母語話者の直感を尊重して異なる音素として数えたが、異音とする考え方もある。ミャンマー語では、介音との合成音素 11 個を子音数に加えた。他言語の介音にはない大きな変化がみとめられるからである。

*6　英語は北米、スペイン語はスペイン王国、ポルトガル語はブラジル連邦共和国の標準語が開発対象である。

*7　話者数の国別・地域別比率である。原則として L1 話者内で集計した。主に話されている国が一つの場合は、その国内での比率である。主に話されている国が多言語国家で当該言語の母語話者が過半数ではない場合は、国内での他言語を比較対象とした。韓国語は大韓民国の話者数に対する比率である (北朝鮮は含まない)。

*8　音しか表さない文字を「表音文字」、意味も表す文字を「表意文字」とした。

正書法の文字数 *10	正書法における語境界 *11	数字表記 *12	正書法補足（句読法など）*13	ラテン文字公式表記法	語彙表現の性差 *14	数字の読みの規則性 *15	TTS との親和性 *16	諾否疑問文の平叙文との違い *17
33 字	なし	クメール数字 +アラビア数字 ๑๒๓๔๕๖๗๘๙๐ 1234567890 小数点は “ , ” 桁区切りは “ , ”	文末記号は “។/?/!” 章末などより大きな区切りは “ ។ ”	地理部式（2002）	あり	高	高〜中	接尾語句付加 （“ទេ” など） *
44 字	スペース	デーバナーガリ数字 +アラビア数字 १२३४५६७८९० 1234567890	文末記号は “।/?/!” 読点は “ , ” ハイフン使用可	IAST（1894）, ISO 15919（2001）	*	高	高	イントネーションのみ（文字では疑問符のみ）が普通
35 字×大小（2 独自文字：ө, ү）	スペース	アラビア数字 +モンゴル数字 1234567890 ᠐᠑᠒᠓᠔᠕᠖᠗᠘᠙	文頭と固有名詞は大文字，文末記号は “./?/!” ハイフン使用可	なし	なし	高	高	接尾語付加 m: “ɵy”/“yy” f: “ɵү”/“үү”
33 字×大小	スペース	アラビア数字 1234567890 小数点は “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!”，コピュラは “–”	GOST 7.79（2002）, ISO 9 : 1995	あり	中（性格）	高	イントネーションのみ（文字では疑問符のみ）
28 字	スペース（文字により語内で切れるように見える）	アラビア数字 / インド数字 1234567890/ ١٢٣٤٥٦٧٨٩٠	文末記号は “ . / ؟ ”，読点は “ ، ” 口語の正書法は確立されていない ハイフン使用可	ISO 233（1984）	あり	中（性格）	低	接頭語付加（“hal” や “ʔa-”）
27 字×大小（1 独自文字：ß）	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ . ”	文頭と名詞は大文字，文末記号は “./?/!”. ハイフン使用可	-	あり	高（性格）	高	語順
26 字×大小	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!”. 読点は “ , ”. ハイフン使用可	-	あり	高（性複）	高	イントネーションのみ（文字では疑問符のみ）が普通
44 字	スペース	デーバナーガリ数字 +アラビア数字 १२३४५६७८९० 1234567890	文末記号は “।/?/!” 読点は “ , ” ハイフン使用可	IAST（1894）, ISO 15919（2001）	*	高	高	接頭語付加（“ ਜਜ ”） *
33 字×大小（4 独自文字：є, ґ, і, ї）	スペース	アラビア数字 1234567890 小数点は “ , ” 桁区切りは “ , ”	文頭と固有名詞は大文字，文末記号は “./?/!”，コピュラは “-”	GOST 7.79（2002）, ISO 9 : 1995	あり	中（性格）	高	イントネーションのみ（文字では疑問符のみ）が普通

「表語文字」は日本語における定着度が十分でないと思われたため用いていない。数字、単位記号など言語横断的に用いられる表意文字は対象外とした。

*9　アラビア語の書字方向は右から左が基本だが、数字の部分はアラビア数字・インド数字ともに左から右へ書く。

*10　デーバナーガリ文字など Brahmic 系文字の母音記号、アラビア文字の母音記号他各種記号、その他声調・アクセント・鼻音化等を示す付加記号を含まない数である。ベトナム語は英語のアルファベット 26 字から F, J, W, Z を除き独自文字を加えた。ウクライナ語はロシア語のアルファベット 33 字から Ё, Ъ, Ы, Э を除き独自文字を加えた。ウクライナ語は硬音記号「ʼ」を含まない数である。「×大小」は大文字と小文字があることを示す。漢字の数は、日本語は常用漢字 2,136 字+人名用漢字 862 字、中国語は通用規範漢字表の一級字 3,500 字+二級字 3,000 字+三級字 1,605 字とした。

*11　「なし」の言語であっても長い文は適宜読点やスペースで区切る。

*12　字種が複数ある場合はよく使われると思われる方を先に示した。小数点がピリオド、桁区切りがコンマの場合は何も記載していない。モンゴル数字の使用は紙幣などに限定される。アラビア語の「インド数字」は主にアラビア語圏の東側（インドに近い地域）で使われる。

*13　複合語の要素の区切りにハイフンの使用を許す言語は「ハイフン使用可」と記載した。

14　「」は性差のある表現があっても中性の表現で自然に代替可能な言語、「**」は主に聞き手の性差のみの言語を示す。印欧語族言語において中性の表現で代替できない性差は、ほぼ話者自身の属性を示す名詞か話者自身を修飾する形容詞のみで現れる。

*15　「性、複、格」は、修飾する名詞の文法的性、単数／複数、格に依存して変化することを示す。ロシア語、アラビア語、ウクライナ語については、変化自体は規則的であるものの、性・格への依存が特に広範に生じるため評価を「中」とした。

*16　表意文字言語を「低」、表音文字言語で表音性が特に高いものを「高」とし、その他を「中」とした。ただし、表音文字言語でも表音性に明らかな欠落がある場合は「低」とした。たとえば、アラビア語は短母音の有無と種類を表示しない表記法が一般的である。

17　「」はカジュアルなスタイルや場面では「イントネーションのみ（文字では疑問符のみ）」も多い言語である。

読み方が変わるため、低い評価となっている。最後の『諾否疑問文の平叙文との違い』は、「はい」か「いいえ」で答えるタイプの疑問文を同じ内容の平叙文と識別する手がかりが「語尾上げ」などイントネーション以外にあるか、という項目である。ASR・TTS のいずれにも関わるが、前者で課題となることが多い。

3　多言語化の体制

2022 年度の時点で、対象とする 21 言語すべての母語話者が研究室に在籍する。外国語大学／学部を除くと、これだけ多言語の話者がいる研究／開発環境は国内では稀有であろう。彼女ら彼らの担当業務は多岐にわたり（後述）、研究開発のあらゆる段階で欠くべからざる役割を果たしているが、当初からこの体制をとっていたわけではなく、研究開発フェーズの移行とともに段階を経て現在の体制に行き着いた。その変遷の過程と現状を述べる。

まず、最初の段階は国際共同研究の形態であった。各言語における開発をその言語が話される国・地域の研究機関や大学が担当する。ごく初期には各機関同士で一対一の協定を結んでいたが、それは速やかに国際コンソーシアムの枠組へと移行した。これには、新たな機関が加わるごとに関係機関全てと個別協定を結び直す必要がないという利点がある。前身の研究開発における体制が日米独 3 か国の機関による C-STAR（Consortium for Speech Translation Advanced Research, 1991 年～）で始まり、A-STAR（Asian Speech Translation Advanced Research Consortium, 2006 年～）を経て、U-STAR（Universal Speech Translation Advanced Research Consortium, 2010 年～）へと引き継がれた。NICT は 2006 年 4 月から 2016 年 3 月までこれらを事務局として支えた。

国際共同研究の形態は、各言語に依存した人的物的資源が現地で得られやすいという利点があり、また国際研究交流などアピールポイントも多い。しかし、開発が進み、実用システムを組み上げる段階では、各機関における開発ペースの違い等のために、どうしても言語間で不均衡が生じてしまう。そこで、各言語を担当する研究者を NICT で受け入れる形態を一部の言語について取り入れた。主に東南アジア地域の大学・研究機関から招へい研究員・研修員を招き、数言語でほぼ白紙の状態から公開モデルの完成まで漕ぎ着けた。特に、ミャンマー語のシステムは、2015 年 12 月の公開時点では同言語として世界初の音声翻訳アプリであったこともあり、2022 年現在もトップの日本語に迫る利用数を保つ人気を博している [2]。

他機関から研究者を受け入れる形態は、あくまで本人の研鑽・研修が主目的であり、相手機関との間で目的とスケジュールのすり合わせが必要である。より速く広い多言語化の需要に応えるためには、開発の全過程を単一機関の管理下で行うことが理想的である。そこで、研究開発サポート業務を専門とする母語話者の「言語担当者」を研究室内に置く体制へと移行した。これには、開発成果の権利を NICT が保持することが容易になるという利点もあった。言語担当者は国内在住者から募ったが、専門分野が音声や言語であることを必須の条件とはしなかった。これは、言語によっては国内に候補者が僅少であることが予想されたためであったが、研究開発フェーズが言語によらない手法にシフトしていたから可能であったことも事実である。

言語担当者は主に下記のような役割を担い、効率的かつ高品質な多言語化に貢献する。まず、最も基本的な役割は、音声・言語コーパスの品質の向上と維持である。他章で述べたように、機械学習による ASR・TTS では学習に用いる音声・言語データの量と質の両方が性能を左右する。量が少ない場合に質が重要なのは言うまでもないが、量を増やしても質が十分でなければ性能は頭打ちとなる。音声コーパスの質は、コーパス全体としての発話内容の多様性や網羅性、雑音・残響の程度など音響的な健全性のように母語話者でなくても評価可能な部分もあるが、当該言語としての質、すなわち発話の正確性・了解性・自然性、表現の規範性・自然性などの部分は母語話者以外には評価し難い。書き起こしテキストが音声に正しく対応しているか、話者が訛っていないかなどの評価も同様である。

次に、本特集号 **2–2–1** で述べた多言語対訳辞書など翻訳データの品質向上にも言語担当者が貢献する。翻訳自体は外部の翻訳者に依頼するが、対訳辞書の対象は音声・言語コーパスに出現しにくい語彙なので、一般の辞典には掲載されておらず、翻訳の方法が定まっていないものが多い。日本の地名や施設名など固有名詞も多く含まれるため、ローマ字が使える（＝正書法がラテン文字の）言語以外では日本語の転写も問題となる。そこで、NICT の担当者が様々なタイプの語彙に対する翻訳方針を定めた『辞書翻訳ガイドライン』並びに『日本語音写ガイドライン』をあらかじめ作成することで品質の統一を図るとともに事後の評価も担当する。

以上については、言語担当者自らが音声を書き起こしたり、対訳を作成したりすることもあるが、音声・テキストともにデータ量が膨大なため、少量を抜き取って検査した結果をコーパスや翻訳データの作成者に返して改善を促す、という役割が中心である。

最後に、ASR・TTS システムの評価も重要な役割である。システムの更新時にていねいな評価を実施する

とともに、開発メンバーの求めに応じて適宜評価し改善を助ける。このほか、担当言語とその言語が主に使われる国や地域に関する調査やコンサルティングも有効な支援となる。

4 あとがき

現在の研究開発体制は、母語話者が機械よりも優れた聞き手・話し手・言語の使い手であることを暗黙のうちに前提としている。本特集号 **2-2-3** で述べるように、この分野にも機械が人を超える時代がやってくる。体制の変遷は今後も続くものと思われる。

謝辞

表 1 の項目を含む対象言語に関する情報の調査では、大学で各言語を専門とする先生方に助言を仰いだ。併せて、母語話者の言語担当メンバーを紹介いただいた場合も多かった。ここにあらためて謝意を表したい。

【参考文献】

1　The International Phonetic Association, “Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet,” Cambridge University Press, 1999.
2　VoiceTra サポートページ, “累計ダウンロード数と累計発話数,” https://voicetra.nict.go.jp/monitor/ja/VoiceTraSummary.html

加藤 宏明 （かとう ひろあき）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（工学）
音声言語コーパス、聴覚・音声コミュニケーション
【受賞歴】
1995 年　日本音響学会 第 12 回粟屋潔学術奨励賞