Routledge
Taylor & Francis Group

Australian Journal of Linguistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cajl20

Australian English Bilingual Corpus: Automatic forced-alignment accuracy in Russian and English

Ksenia Gnevsheva, Simon Gonzalez & Robert Fromont

To cite this article: Ksenia Gnevsheva, Simon Gonzalez & Robert Fromont (2020) Australian English Bilingual Corpus: Automatic forced-alignment accuracy in Russian and English, Australian Journal of Linguistics, 40:2, 182-193, DOI: 10.1080/07268602.2020.1737507

To link to this article: https://doi.org/10.1080/07268602.2020.1737507

AUSTRALIAN JOURNAL OF LINGUISTICS
2020, VOL. 40, NO. 2, 182–193
https://doi.org/10.1080/07268602.2020.1737507

Australian English Bilingual Corpus: Automatic forced-alignment accuracy in Russian and English

Ksenia Gnevsheva, Simon Gonzalez and Robert Fromont

Australian National University, Australian National University and University of Canterbury

ABSTRACT

This paper introduces the Australian English Bilingual Corpus, a Russian–English spoken corpus, and uses it for a comparison of automatic time alignment between two different languages. Automatic forced alignment is gaining popularity in corpus research as it allows for time-efficient processing of phonetic information. The Language, Brain and Behaviour: Corpus Analysis Tool is one aligner which compares well with others in terms of alignment accuracy. Most of the forced-alignment work has been done with different varieties of English. This paper compares alignment accuracy between Russian and English and discusses aligner settings and data characteristics that affect it. The results suggest higher alignment accuracy for English than Russian. For Russian, alignment accuracy improves with stress specification; that is, when stressed and unstressed vowels are treated as separate categories.

ARTICLE HISTORY
Accepted 18 February 2020

KEYWORDS
Forced alignment; alignment accuracy; bilingual corpus; Australian English; Russian

1. Introduction

Recent advances in technology have made acoustic analysis of large spoken corpora more efficient. Forced alignment allows researchers to automatically time-align audio-recordings and orthographic transcriptions at the level of the segment. This speeds up analysis and can increase the number of tokens from 300 to 9,000 (Labov et al. 2013). Some of the most widely used forced aligners are Forced Alignment and Vowel Extraction program suite (FAVE; Rosenfelder et al. 2011), Language, Brain and Behaviour: Corpus Analysis Tool (LaBB-CAT; Fromont & Hay 2012), Montreal Forced Aligner (MFA; McAuliffe et al. 2017), and Munich Automatic Segmentation System (MAUS; Kisler et al. 2012). Because of the differences in algorithm (HTK (University of Cambridge 2014) vs Kaldi (Povey et al. 2011)) and model-training methods (pre-trained vs train/align protocols), the existing aligners vary in their alignment accuracy. For example, when comparing the alignment of Australian English by four of the most commonly used aligners, LaBB-CAT shows more accurate alignment than FAVE and MAUS, and similar accuracy to MFA (Gonzalez et al. forthcoming).

LaBB-CAT was originally developed for the Origins of New Zealand English corpus (Fromont & Hay 2012) and, consequently, used most for research on New Zealand English (e.g. Hay & Foulkes 2016). However, it can be paired with any dictionary and

CONTACT Ksenia Gnevsheva ksenia.gnevsheva@anu.edu.au

does not require a pre-trained acoustic model (LaBB-CAT uses a 'train and align' protocol; that is a model is trained on the data provided). This allows it to be used for different linguistic varieties (but see DiCanio et al. (2013), Kempton et al. (2011) and Kurtic et al. (2012) for examples of languages aligned with acoustic models trained on a different language). LaBB-CAT has been successfully used for England English (Clark & Watson 2016), Scottish English (Rathcke & Stuart-Smith 2016), Australian English (Docherty et al. 2015) and second language English (Gnevsheva 2015a, 2015b) corpora. LaBB-CAT, and other aligners, are more rarely used for forced alignment of other languages (but see King et al. (2011) for Māori) despite the urgent need for automation of data processing for linguistic research on smaller languages (Barth et al. 2020).

There are a number of data characteristics and LaBB-CAT settings that can affect the quality of alignment. Gonzalez et al. (forthcoming) find that acoustic models trained on spontaneous speech outperform those trained on scripted speech type and vowel onset boundaries are more accurate than offset boundaries (which is also found to be the case for other aligners). These findings are in line with Fromont & Watson (2016) who investigated the effect of pause marking, overlapping noise, number of speakers, different sampling rates, using pre-trained vs 'train and align' models, and amount of data on forced-alignment accuracy. They find that the 'train and align' approach is maximized when (i) enough data are available – at least 5 min of speech per speaker, and (ii) the acoustic model is trained on spontaneous speech. In general, forced alignment was found to be quite robust in the presence of other speakers, overlapping noise, absence of pause marking and downsampling. In addition, the study used speakers from three different English varieties (New Zealand, US and UK English data), and although some differences were found (e.g. in the effect of downsampling), the effect of linguistic variety was not the main focus of investigation. However, in the context of increasing application of automatic alignment in language documentation work, it is important to know how alignment accuracy differs across languages and what aligner or data manipulations can be done to improve the result.

In sum, as most work is done on a handful of large languages, technology facilitating research on smaller languages is of immense importance. Automatic alignment can help to bridge the gap between major and minor languages in corpus compilation, but it is as yet unclear whether alignment quality would differ across languages, and a better understanding of factors affecting it is needed. The aim of this paper is to compare LaBB-CAT's alignment accuracy across two languages, Russian and English. We begin by introducing the Australian English Bilingual Corpus (AusEBC), which currently houses the data, and then use a representative subset to investigate alignment accuracy in the two languages. We find the alignment of English data to be more robust and attribute it to differences in the two languages' vowel systems. We also find that Russian alignment can be improved by treating stressed and unstressed vowels as separate categories.

2. The Australian English Bilingual Corpus

Most of the existing spoken corpora contain data from a single language. Corpora containing audio-recordings in both of a bilingual's languages are quite rare (but see King et al. (2011) for English and Māori and Travis & Torres Cacoullos (2013) for English and Spanish); however, they provide rich data for understanding bilingual processes, be it

language acquisition or language attrition. Moreover, the existing corpora usually record participants in a single style (most often an interview). The Australian English Bilingual Corpus was created for the study of sociolinguistic variation in bilingual speakers, and to this effect it includes audio-recordings of bilinguals using their two languages in several styles.

The corpus currently comprises speech data from Russian–English bilinguals, but there are plans to expand it to other languages. Russian was chosen as the starting point as it has not received much attention in the Australian context to date despite 85,657 Australians claiming Russian heritage and 50,314 people speaking Russian at home (censusdata.abs.gov.au, 2016). Most of Russian Australians live in Melbourne and Sydney; data collection for this corpus was conducted in Melbourne.

Sixteen Russian–English bilinguals comprise the corpus: five males and five females who learned English as a foreign language and moved to Australia after the age of 18 (generation 1 migrants, Gen 1) and five females and one male who were born in Australia (or moved there before the age of five) and learned Russian as their heritage language at home and through weekend Russian community school (generation 2 migrants, Gen 2). On average, the Gen 1 participants had lived in Australia for 4 years; none of the Gen 2 participants had lived in Russia (except for one female participant who was born in the USSR and lived there for the first 2 years of her life). When asked to self-report their ability in both languages on a 1–5 Likert scale (1 = poor to 5 = native-like), Gen 1 speakers on average rated themselves as 3.55 and 4.75 in English and Russian respectively and Gen 2 speakers as 4.25 and 2.5. In fact two of the six Gen 2 participants' reading passage recordings were discarded because of high disfluency (see below). Eight participants reported proficiency in other languages (six of Gen 1 and two of Gen 2). The participants' ethnic orientation was quantified by having them complete a questionnaire which asked about their language use, cultural heritage and ethnic identification (following Hoffman & Walker 2010). The responses were transferred to a three-point scale and averaged for each participant, so that a higher number means higher engagement in the ethnic group. Gen 1 participants averaged an ethnic orientation score of 2.26 and Gen 2 participants 1.89. The participants were between the ages of 18 and 40 (mean 29) at the time of data collection and reported regular use of both languages in their daily lives. All had had some tertiary education: 12 had at least a Bachelor's degree (with five post-graduate degrees) and three of the remaining four were university students at the time of the study.

An Hn5 Zoom audio-recorder and a Samson head-mounted microphone were used for data collection. The participants were audio-recorded speaking their two languages in three spontaneous and controlled production tasks in a quiet space. They: (1) self-recorded two natural conversations with friends; (2) participated in two sociolinguistic interviews with a male Russian–English bilingual research assistant (Gen 2[1]); and (3) read two standardized reading passages (a combination of the North Wind and the Sun, Grandfather, and Rainbow passages) in Russian and English (see Appendix A). This resulted in about 1 h of recording per speaker per language (a minimum of 5 min of participant speech in each of the three styles). The audio-recordings were anonymized by introducing noise over parts mentioning people's names or addresses; references to large areas where identification of

[1]See Travis & Torres Cacoullos (2013) and references therein for the argument that data should be collected by in-group members.

the speaker based on such information would be improbable (e.g. Melbourne or Moscow) were retained. Biographical and ethnic orientation information was gathered through a questionnaire after the production tasks (ensuring that the data were not influenced by participants' analysis of their behaviour which may be prompted by the questionnaire).

Spoken corpora that include several speaking styles are quite rare, and most phonetic work is based on reading tasks: wordlists or passages. In the Australian context in particular, most research on production has employed more controlled methods, such as reading lists (e.g. Cox 2006), allowing for greater control over phonological environment and recording quality. The sociolinguistic interview has also been one of the preferred data collection methods for linguists because it arguably allows the researcher to access the most unmonitored speech by engaging the participant in a conversation that is of interest to them (Labov 1972). During this data collection the participants were asked about their sociodemographic information, family, cultural identification, language use and diversity. While these two tasks are quite common in variationist sociolinguistic studies, self-recording participants in their everyday communication is much rarer (but see Gnevsheva 2015a; 2015b; Sharma 2011), despite being more naturalistic and providing more ecological validity to the data-collection procedure. Collecting several types of recording from the same participants allows the researcher to investigate the speakers' stylistic repertoire.

During corpus setup, the anonymized audio-recordings were manually segmented at the level of utterance and orthographically transcribed in ELAN (2017). The audio-recordings and transcription files were uploaded into LaBB-CAT. A 'phonemes' layer was configured by automatically generating phonemic transcription from orthographic transcription; CELEX (Baayen et al. 1995) was set up as the main dictionary for English, and a Russian dictionary (from MFA) was set up as an auxiliary dictionary for Russian. After phonemic transcriptions were created, automatic alignment was done using HTK (University of Cambridge 2014). The whole corpus comprises over 200,000 words.

English and Russian are both stress-timed languages with significant vowel reduction in unstressed positions. English was aligned as a system with 19 vowels with no stress specification, but only the alignment of 11 monophthongs is analysed here. Russian was aligned two times for comparison: as a six-vowel system without stress specification and as a 12-vowel system when stress is specified, treating stressed and unstressed variants of the same vowel as distinct categories in the dictionary.

After such alignment is done and a 'segments' layer is created, acoustic information can be accessed and exported automatically through direct integration with Praat (Boersma & Weenink 2018) or in any analysis software of choice after exporting Praat TextGrids. To illustrate this, we extracted and plotted the first two formant frequencies for monophthongs produced by a single female speaker (Gen 1) in a reading passage in English and Russian (Figure 1 left and right panels respectively).

For example, such data extracted from the corpus have been used for an analysis of bilingual speakers' style-shifting in English (Gnevsheva 2015b).

3. Alignment accuracy

Having similar data produced by the same speakers in two different languages lends itself well to a comparison of alignment accuracy across languages. Comparing bilingual speakers with similar proficiency in the two languages helps to minimize the effect of fluency on

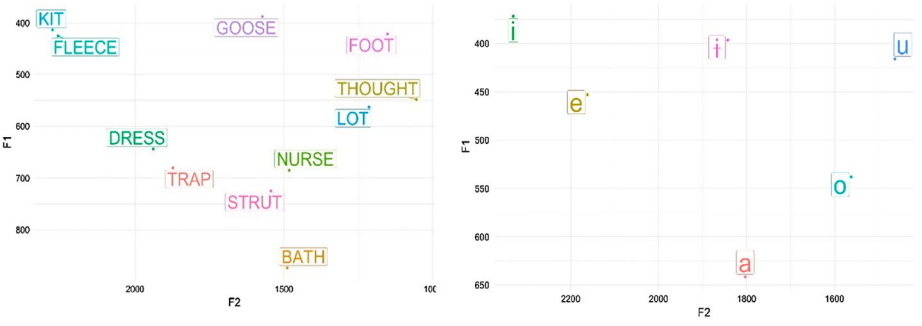

Figure 1 Gen 1 speaker's English (left) and Russian (right) vowel spaces

alignment accuracy. In the analysis of alignment accuracy, we focus on vowel boundary displacement in the reading passage by four speakers from AusEBC (two Gen 1 and two Gen 2 speakers most fluent in both languages but dominant in Russian and English respectively; one male and one female in each pair). Boundary displacement is the time difference in milliseconds (ms) between the human benchmark and the automatic boundary placed by LaBB-CAT. Low displacements represent more accurate alignment, whereas high displacements represent less accurate alignment (see Figure 2).

To make the data as comparable as possible, alignment accuracy was calculated for the same part of the reading passage for all the speakers (about 100 words; a total of about 1,500 vowel tokens across four speakers; Table 1). A human benchmark for assessing alignment accuracy was established by extracting Praat TextGrids and manually correcting vowel boundaries by one of the authors (a trained phonetician, native speaker of Russian and near-native speaker of English).

Table 2 details boundary displacement for the onset and offset of stressed and unstressed vowels in the two languages. The positive skew in the distributions is attested by the difference between means and medians and is explained by a few outliers with very high boundary displacement. The median boundary displacement for English vowels compares favourably with the 20 ms error threshold (Cosi et al. 1991). English boundary displacement is noticeably smaller compared to Russian. There are some numerical differences

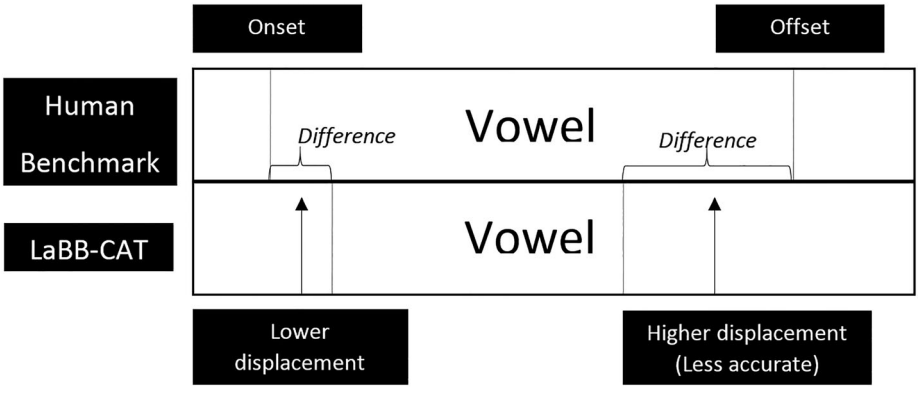

Figure 2 Boundary displacement

Table 1 Number of vowel tokens in the dataset

Vowels	Number of tokens
English stressed	498
English unstressed	133
Russian stressed	406
Russian unstressed	477
Total	**1,514**

Table 2 Boundary displacement for English and Russian in ms

Language	Stress	Position	Mean	Median	SD
English	Stressed	Onset	26.1	11.8	56.2
		Offset	19.4	7.51	56.5
	Unstressed	Onset	21.2	9.54	51.9
		Offset	24.3	5.11	67.4
Russian	Stressed	Onset	66.9	17.4	192.
		Offset	59.9	18.	183.
	Unstressed	Onset	74.1	18.8	225.
		Offset	86.8	18.9	251.

between onsets and offsets as well as stressed and unstressed vowels, but any trends are less readily identifiable by looking at the raw numbers, and statistical modelling is needed for significance testing.

4. Statistical analysis

Statistical analysis for significance of differences in boundary displacement was performed by fitting mixed effects models (Baayen et al. 2008) in R (R Core Team 2017). To test for significant differences between the two languages, we ran a model with boundary displacement as the response variable, language (English vs Russian) and position (onset vs offset) as the fixed factors, and speaker as a random factor. We also tested the interaction between language and position.

We ran another model on the Russian data only to test for significant differences between alignment with stress specification (as a 12-vowel system) and alignment with no stress specification (as a six-vowel system). Boundary displacement for a given vowel was the response variable; data source (stress and no stress specification dictionary) and stress (whether the vowel was lexically stressed or not) were the fixed factors. Speaker was a random factor. We also tested the interaction between data source and stress.

Both models were pruned to exclude non-significant effects. Final model output (Tables 3 and 4) was created using the sjPlot package (Lüdecke 2018). Larger values in the Estimates column mean larger boundary displacement and, therefore, lower alignment accuracy.

5. Factors affecting alignment accuracy

When comparing alignment accuracy in the two languages, boundary displacement was higher for Russian than English when stress was not specified ($p < 0.001$; Table 3; Figure 3).

Table 3 Mixed effects model output comparing English and Russian

Predictors	Estimates	CI	*p*
(Intercept)	22.21	−21.57–65.99	0.320
language_Russian	67.05	52.52–81.57	<0.001
Random effects			
σ^2	40,301.79		
$T_{00\ speaker}$	1,867.64		
$ICC_{speaker}$	0.04		
Observations	3,026		
Marginal R^2/Conditional R^2	0.025/0.068		

Table 4 Mixed effects model output comparing stressed and unstressed vowels in Russian

Predictors	Estimates	CI	*p*
(Intercept)	0.07	0.01–0.13	0.026
Data_sourceStress_specification	−0.01	−0.03–0.00	0.066
StressUnstressed	0.03	0.01–0.05	<0.001
Random effects			
σ^2	0.05		
$T_{00\ speaker}$	0.00		
$ICC_{speaker}$	0.07		
Observations	3,532		
Marginal R^2/Conditional R^2	0.005/0.073		

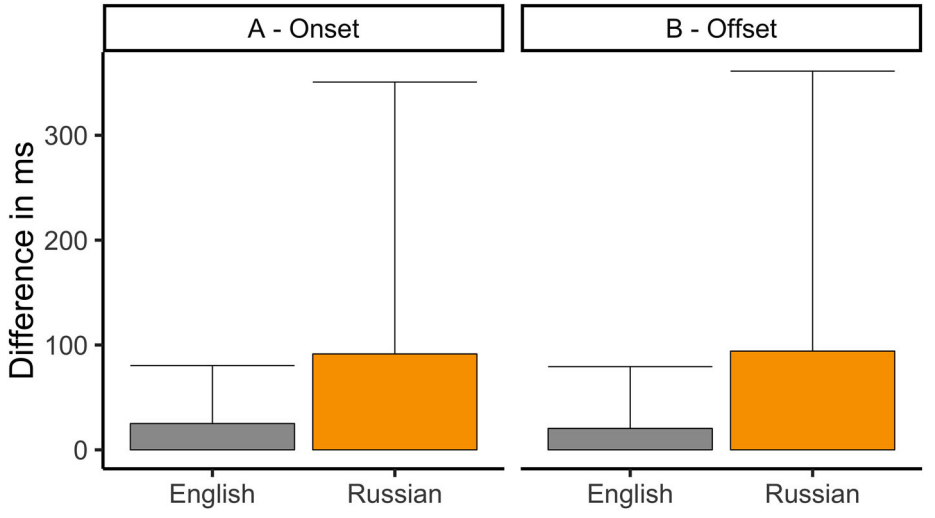

Figure 3 Boundary displacement by language

The difference in boundary displacement between the languages was numerically smaller for vowel onsets than vowel offsets, but not significant ($p = 0.684$). The pattern was the same across languages (no significant interaction; $p = 0.625$), so vowel onsets and offsets are combined for the following analysis of the stress effect.

The difference in alignment accuracy between two stress-timed languages can be explained through the differences in the size of the vowel inventories. Systems with fewer vowels (Russian with six vowels in comparison to English with 19) may have

more variation within categories. This would necessitate more generalized, less discriminative acoustic models to be trained, which would result in higher boundary displacement in alignment. Additionally, the existence of two degrees of reduction in Russian unstressed vowels (Yanushevskaya & Bunčić 2015) introduces even more within-category variation. To investigate this, we tested whether stress specification would make a difference for alignment accuracy in Russian.

In a comparison of alignment accuracy with and without stress specification in Russian (as a 12- or 6-vowel system respectively), boundary displacements were significantly shorter for stressed vowels ($p < 0.001$; Table 4; Figure 4).

Stressed vowels performed better than unstressed ones, perhaps because of differences in length and the amount of within-category variation. Separating stressed and unstressed vowels in Russian alignment improved overall alignment accuracy ($p = 0.066$), possibly due to reduced within-category variation. Stress specification improved alignment for both stressed and unstressed vowels (interaction was not significant, $p = 0.63$).

Giving HTK more categories to discriminate apparently leads to more accurate acoustic models, presumably due to reduced within-category variation. HTK codes the acoustic signal as 12 Mel Frequency Cepstral Coefficients (MFCCs), which are then used as the input data into the training process. MFCCs do not straightforwardly relate to phonetic features commonly used by phoneticians, and so it can be difficult to interpret the exact features that HTK might be using to tease apart and better discriminate the stressed and unstressed vowels. However, in some cases, the acoustic differences between them are easy to see just in terms of formant values. For example, Figure 5 shows a scatter plot of normalized F1 and F2 values of tokens of /a/, with the stressed (A) and unstressed (a) versions of the vowel plotted in different colours.

Although there is substantial overlap, these two categories are clearly different even in terms of just these two dimensions. Giving HTK two categories instead of one for this vowel will allow it (given enough data) to arrive at two more specific acoustic models, rather than one general one, facilitating higher alignment accuracy.

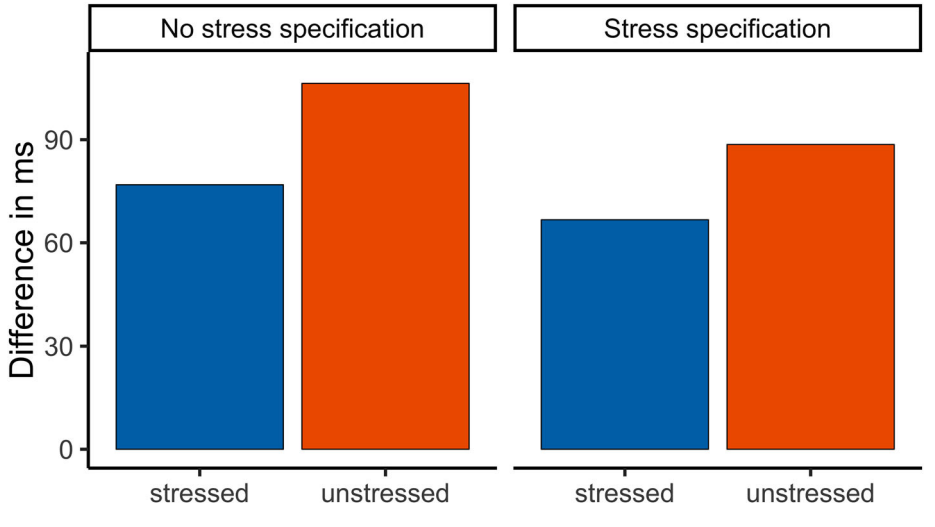

Figure 4 Boundary displacement in Russian

Figure 5 Stressed and unstressed /a/ plots: a – unstressed, A – stressed

6. Conclusions

This study analysed forced-alignment accuracy in two languages, English and Russian, using comparable data from the same speakers. The results suggest that alignment accuracy may vary across languages: alignment accuracy was higher for English than Russian. Researchers should be aware of such differences and not assume a certain degree of alignment accuracy in one language based on another language's data.

Moreover, it is beneficial to know what data characteristics contribute to higher alignment accuracy. For example, in this study alignment accuracy was higher for stressed vowels compared to unstressed ones, and separating stressed and unstressed vowels in Russian improved overall alignment. Stress specification improved alignment for both stressed and unstressed vowels. By specifying stress during alignment and choosing to analyse stressed vowels only, the researcher will arrive at more reliable measurements.

To sum up, it is important to know what language characteristics affect alignment accuracy in order to optimize forced alignment. Based on the results from this study, we provide the following recommendations for improving the accuracy of automatic alignment: (1) stress specification during forced alignment improves alignment accuracy of vowels in stress-timed languages; and (2) automatic alignment of stressed vowels is more robust making them a more reliable object of analysis.

Naturally, these findings come from two languages. Further comparisons need to be done for languages differing in their characteristics, e.g. the size of their vowel inventories

and presence/absence of vowel reduction. Especially interesting for that would be Australian languages like Walpiri, which only has three vowels.

Acknowledgements

This research was supported by the ARC Centre of Excellence for the Dynamics of Language's Transdisciplinary & Innovation Grant scheme (TIG662017 and TIG952018). We are grateful to the participants who volunteered their time and Ben Volchok for help with data collection and transcription. We would like to thank Catherine Travis and James Grama, the audience of the 2018 Indiana University PhonFest, two anonymous reviewers, and the Journal Editor for feedback on earlier versions of this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the ARC Centre of Excellence for the Dynamics of Language, Australian Research Council: [Grant Number TIG662017; TIG952018].

Notes on contributors

Ksenia Gnevsheva is a Lecturer in Linguistics in the School of Literature, Languages and Linguistics at the ANU. She holds a PhD in Linguistics from the University of Canterbury (New Zealand). Her main research interest lies at the intersection of sociolinguistics and second language acquisition. Her current work focuses on sociolinguistic variation in bilingual speakers in production and perception.

Simon Gonzalez is a Postdoctoral Fellow at the Centre of Excellence for the Dynamics of Language. His research focuses on acoustic phonetics, empowered by computational tools. After finishing his PhD in English Phonology (Australian English) at the University of Newcastle, he worked as a Research Assistant at Griffith University analysing West Australian English (ARC-funded, led by Gerard Docherty). He develops computational tools (scripts and online apps) for more efficient and practical analysis/visualization of phonetic and phonological phenomena.

Robert Fromont is a Software Programmer affiliated with the New Zealand Institute of Language, Brain and Behaviour. He is the developer of the Language, Brain and Behaviour: Corpus Analysis Tool.

References

Baayen, R. H., D. J. Davidson & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412.

Baayen, R. H., R. Piepenbrock & L. Gulikers. 1995. *The CELEX lexical database (CD-ROM)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Barth, D., J. Grama, S. Gonzalez & C. Travis. 2020. Using forced alignment for sociophonetic research on a minority language. *University of Pennsylvania Working Papers in Linguistics* 25(2). 2.

Boersma, P. & D. Weenink. 2018. Praat: Doing phonetics by computer (Version 6.0.39). http://www.praat.org/.

Clark, L. & K. Watson. 2016. Phonological leveling, diffusion, and divergence: /t/lenition in Liverpool and its hinterland. *Language Variation and Change* 28(1). 31–62.

Cosi, P., D. Falavigna & M. Omologo. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In Proceedings of the *Second European conference on speech communication and technology*, 693–696.

Cox, F. 2006. The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics* 26(2). 147–179.

DiCanio, C., H. Nam, D. H. Whalen, H. T. Bunnell, J. D. Amith & R. Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134(3). 2235–2246.

Docherty, G., S. Gonzalez & N. Mitchell. 2015. Static vs. dynamic perspectives on the realization of vowel nuclei in West Australian English. *In Proceedings of the 18th international congress of phonetic sciences*.

ELAN (Version 5.0.0-beta) [Computer software]. 2017, April 18. Nijmegen: Max Planck Institute for Psycholinguistics. https://tla.mpi.nl/tools/tla-tools/elan/.

Fromont, R. & J. Hay. 2012. LaBB-CAT: An annotation store. In P. Cook & S. Nowson (eds.), *Proceedings of Australaisian language technology association workshop*, 113–117. Dunedin, New Zealand.

Fromont, R. & K. Watson. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora* 11(3). 401–431.

Gnevsheva, K. 2015a. Acoustic analysis in Accent of Non-Native English (ANNE) corpus. *International Journal of Learner Corpus Research* 1(2). 256–267.

Gnevsheva, K. 2015b. Style-shifting and intra-speaker variation in the vowel production of non-native speakers of New Zealand English. *Journal of Second Language Pronunciation* 1(2). 135–156.

Gonzalez, S., J. Grama & C. E. Travis. forthcoming. Comparing the performance of major forced aligners used in sociophonetic research. *Linguistics Vanguard*.

Hay, J. & P. Foulkes. 2016. The evolution of medial /t/ over real and remembered time. *Language* 92(2). 298–330.

Hoffman, M. F. & J. A. Walker. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change* 22(1). 37–67.

Kempton, T., R. K. Moore & T. Hain. 2011. Cross-language phone recognition when the target language phoneme inventory is not known. *Proceedings of the 12th annual conference of the international speech communication association*, 3165–3168. Florence, Italy.

King J., M. Maclagan, R. Harlow, P. Keegan & C. Watson. 2011. The MAONZE corpus: Transcribing and analysing Māori speech. *New Zealand Studies in Applied Linguistics* 17(1). 32–48.

Kisler, T., F. Schiel & H. Sloetjes. 2012. *Signal processing via web services: The use case WebMAUS*. Paper presented at Digital Humanities Conference, Hamburg, Germany.

Kurtic, E., B. Wells, G. J. Brown, T. Kempton & A. Aker. 2012. A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English. *Proceedings of the eighth international conference on language resources and evaluation*, 1323–1327. Istanbul, Turkey.

Labov, W. 1972. Some principles of linguistic methodology. *Language in Society* 1(1). 97–120.

Labov, W., I. Rosenfelder & J. Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation reversal, and reanalysis. *Language* 89(1). 30–65.

Lüdecke, D. 2018. _sjPlot: Data visualization for statistics in social science_. doi:10.5281/zenodo.1308157 (URL: http://doi.org/10.5281/zenodo.1308157), R package version 2.5.0, <URL: https://CRAN.R-project.org/package=sjPlot>.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner & M. Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the 18th conference of the international speech communication association*, 498–502. Stockholm: ISCA.

Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, Petr Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer & K. Vesely. 2011. The Kaldi Speech Recognition Toolkit. *IEEE automatic speech recognition and understanding workshop*, 4. Hawai'i, USA.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rathcke, T., & J. Stuart-Smith. 2016. On the tail of the Scottish Vowel Length Rule in Glasgow. *Language and Speech* 59(3). 404–430.

Rosenfelder, I., J. Fruehwald, K. Evanini & J. Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite.
Sharma, D. 2011. Style repertoire and social change in British Asian English. *Journal of Sociolinguistics* 15(4). 464–492.
Travis, C., & R. Torres Cacoullos. 2013. Making voices count: Corpus compilation in bilingual communities. *Australian Journal of Linguistics* 33(2). 170–194.
University of Cambridge. 2014. HTK. http://htk.eng.cam.ac.uk/ (1 December 2017).
Yanushevskaya, I., & D. Bunčić. 2015. Russian. *Journal of the International Phonetic Association* 45(2). 221–228.

Appendix A

The North Wind and the Sun

The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

The grandfather passage

You wished to know all about my grandfather. Well, he is nearly ninety-three years old; he dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a trifle. Twice each day he plays skilfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, 'Banana oil!' Grandfather likes to be modern in his language.

The rainbow passage

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Others have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbows. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the drops, and the width of the coloured band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of super-imposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.