LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 13 (2019), pp. 281–299
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24869
Revised Version Received: 25 Apr 2019

Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language

Caroline Jones
MARCS Institute for Brain, Behaviour & Development,
Western Sydney University

Weicong Li
MARCS Institute for Brain, Behaviour & Development,
Western Sydney University

Andre Almeida
MARCS Institute for Brain, Behaviour & Development,
Western Sydney University

Amit German
MARCS Institute for Brain, Behaviour & Development,
Western Sydney University

Speech technology is transforming language documentation; acoustic models trained on "small" languages are now technically feasible. At the same time, forced alignment built for major world languages has matured and now offers ease of use through web interfaces requiring low technical expertise. This paper provides an updated and detailed evaluation of *cross-linguistic forced alignment*, the approach of using forced aligners untrained on the target language. We compare two options within MAUS (Munich Automatic Segmentation System): language-independent mode vs major world language system (here, Italian) on the one dataset, a comparison that has not previously been reported. The dataset comes from a corpus of adult conversational speech in Kriol, an English-based creole of northern Australia. The results of using MAUS Italian were better than those of using the language-independent mode and those in previous studies: the agreement rate at 20 ms was 72.1% at vowel onset and 57.2% at vowel offset. With completely misaligned tokens excluded, the overall agreement rate rose to 69.2% at 20 ms and over 90% at 50 ms. Most errors in the output SAMPA (Speech Assessment Methods Phonetic Alphabet) labels were resolvable with simple text replacements. These results offer updated benchmark data for an untrained, late-model forced alignment system.

E-ISSN 1934-5275

1. Introduction Rapid advances in speech technology are changing the face of workflows in language documentation. It is now possible to train acoustic models for forced alignment of transcript to audio based on relatively small datasets (Johnson et al. 2018) and even to perform a first-pass automatic transcription (Adams 2018). These advances hold the promise of resolving issues of scale which have plagued documentation work in "small" languages, i.e. under-resourced and/or under-documented. For example, the fact that forced alignment systems for major world languages had been developed with large datasets (requiring considerable investments of time, specialist labour, and funding, in addition to well-specified lexicons and pronunciation rules) suggested that forced alignment systems designed specifically for low-resource languages would remain relatively unachievable. In that context, a handful of evaluation studies (Kempton et al. 2011; Kurtic et al. 2012; DiCanio et al. 2013; Strunk et al. 2014; Kempton 2017) were done with forced aligners, dating from the late 2000s and early 2010s, to find out if it was viable to use cross-linguistic forced alignment (i.e., the use of an aligner developed for a major world language applied to a "small" language). The resulting agreement rates (with a human labeller as "gold standard", within 20 ms) were 41–66%, with values of 41–49% for conversational data – not strong results, as noted by Johnson et al. (2018), particularly when compared with rates over 80% for aligners used with major world language data.

As a general approach, however, cross-linguistic forced alignment remains attractive where there is a lack of technical resourcing, and if the system is user-friendly. Much of the small cross-linguistic forced alignment literature is based on forced alignment systems that are now quite old. Newer systems like MAUS (Munich Automatic Segmentation System) have been evaluated for "small" languages in only one previous published paper (Strunk et al. 2014). In this paper we offer an updated evaluation of cross-linguistic forced alignment at a point in time when systems like MAUS are mature, offering very easy-to-use web interfaces (e.g., WebMAUS; Kisler et al. 2017), unlike systems that require training (e.g., for anything other than American English, such as Prosodylab-Aligner, or Montreal Forced Aligner, using Kaldi). Anecdotally many colleagues have trialled or used MAUS in the language-independent (*sampa*) mode, some having difficulties when using conversational data. In this paper we offer a careful comparison of the results of two MAUS options applied to the same conversational dataset: MAUS in language-independent mode versus major world language (here, Italian). To our knowledge this comparison has not been made previously. We offer, in addition, more detailed evaluation data than has previously been typical in the small reports of cross-linguistic forced alignment.

The dataset involves a low-resource language, Kriol, an English-based creole language of northern Australia in its local variety spoken at Barunga Community, Northern Territory. The dataset comprises spontaneous speech in conversational format by young adult native speakers. It is hoped that this careful comparison of forced alignment options might be helpful for working linguists interested in trying forced alignment for themselves, in addition to providing updated benchmark evaluation data for mature forced alignment systems in untrained mode.

2. Background

2.1 Kriol Phonology Project[1] The dataset for the current study is a subset of a new corpus of north Australian Kriol created in 2014–2017 with speakers at Barunga, a remote Aboriginal community near the town of Katherine, Northern Territory, Australia. North Australian Kriol is an English-based creole language spoken by approximately 30,000 speakers across northern Australia, typically by Aboriginal adults and children as their everyday home language, in remote and regional areas where the traditional languages tend to be spoken fluently only by the very most elderly members of the community. A typical example of a sentence in the Barunga variety of north Australian Kriol is provided in (1). This example illustrates how the words are historically derived from English words (e.g., *rait* > right, *bas* > bus), but sometimes differ phonologically reflecting traditional language phonologies (e.g., medial trill in *garra* > gottem, lack of /h/ in *im* > him), and how the orthography is a regular spelling system (again like the traditional languages of the area).

(1) *o im rait ai gobek garra bas na*
oh 3SG right 1SG go.back by bus DM
'Don't worry, I'll go back on the bus.'
[CJ2-059-01_84521_86508. wav]

The subset of the corpus for the current study comprises spontaneous speech (conversations about everyday topics), produced by five young adult native speakers of the local Barunga variety of Kriol. Speakers were audio-recorded using a linear PCM recorder (Olympus, Australia, LS-14 model) with a Rode lapel microphone, at 44.1 kHz, 16-bit, in quiet outdoor field conditions, and in conversation with a familiar local Aboriginal age-peer from the same community. The corpus was orthographically transcribed in ELAN (version 4.9.4)[2] by non-native speaker linguists familiar with the language and checked with native speakers. The recordings used in this study (total duration 132 minutes) in annotated form comprise a total 10,395 transcribed words. For the analysis of vowel acoustics, an early interest in our work, the focus was on unreduced vowels in prominent syllables in content words. A total of 1050 vowel tokens (647 short vowels, 186 long vowels, and 217 diphthongs) meeting these criteria were used for the study.

2.2 WebMAUS: two workflow options WebMAUS is the web-based version of the MAUS. The WebMAUS system is structured into several modules which can each perform steps in the workflow. One option for a "small" language is to use the

[1]The Kriol corpus was recorded by Tiarnah Ahfat, Delvean Ahfat, and Anita Painter. The authors would like to thank Sarah Cutfield for assistance with transcription, and undergraduate student summer interns Thomas Batchelor, Jessica Chin, Adrienne Grant, and Natasha Hollamby for assistance with editing segment boundaries. The research was supported by Australian Research Council Future Fellowship FT120100777 and CI funding from Australian Research Council Centre of Excellence for the Dynamics of Language CE140100041.

[2]https://tla.mpi.nl/tools/tla-tools/elan/.

WebMAUS General language-independent system. This is possible if annotation in SAMPA (Speech Assessment Methods Phonetic Alphabet) already exists. It is also possible if a SAMPA annotation can be simply created using a script that applies basic grapheme-to-phoneme (G2P) rules and formats annotation in BPF format (Bavarian Archive for Speech Signals Partitur Format), which has a few advantages over other existing formats for transcripts, labelling, and segmentation (Schiel et al. 1997). This system includes acoustic models and SAMPA symbols from all the languages for which MAUS has developed trained systems. This language-independent approach seems to have gathered the most attention from fieldworkers and other researchers on "small" languages.

The alternative is to use several modules which are language-specific to major world languages (e.g., English, Spanish, and Italian). Starting with an original orthographic transcription (e.g., in plain text, ELAN, Transcriber, or EMU) plus audio file, it is possible to prepare conversational files using the Chunk Preparation service. This should make for better alignment for conversational data. The output of Chunk Preparation is a BPF file which incorporates utterance start and end points, the orthographic transcription of each utterance, a tokenised version of ordered words, and their conversions into SAMPA. After Chunk Preparation, the BPF file can be used as input to WebMAUS General with the selection of a major world language for the acoustic models.[3]

Chunk Preparation is only available for major world languages, and some of the MAUS major world language systems have been trained intensively on spontaneous, relatively noisy data. This means that opting to use an existing system for a major world language may bring particular benefits if the "small" language data is conversational in form and has been recorded in other than lab conditions. On the downside it can be anticipated that the G2P rules applied for the world language (in the Chunk Preparation mode) may result in alignment errors when the MAUS General goes to align the transcription with the "small" language data. The more different the "small" language is from the world language, the more errors in orthography, phonology, and phonetics.

Both options – language-independent and major world language – seem to have advantages and disadvantages. Which option works best for a "small" language dataset is currently an open question, and to an extent, the question will always be dataset-specific. In any case, it is clear that the comprehensiveness and the accuracy of transcription can be major factors in the success of forced alignment (Strunk et al. 2014). In this paper we keep the transcription constant, in a bid to explore the relative merits of different forced alignment options for "small" languages.

3. Evaluating MAUS options for Barunga Kriol The major world language MAUS system which we chose to apply was Italian. The reasons we chose Italian (at the suggestion of Florian Schiel, MAUS developer) were three. First, Italian has a relatively phonetic (i.e., transparent) or regular orthography like Kriol (unlike English).

[3]Note that it is not currently an option within MAUS to write language-specific G2P rules (for the "small" language) and then combine them within a major-language alignment model.

Second, Italian, with seven monophthongs, has a similar vowel system to Kriol. This variety of Kriol has a system of five short vowels /ɪ, ɛ, ɐ, ɔ, ʊ/ (with a possible 6th /æ/ vowel), five rarer long vowels /iː, ɜː, uː, oː, ɐː/, and five diphthongs /eɪ, ɐʊ, ɐɪ, oʊ, oɪ/ (Jones et al. 2017). Third, and perhaps most importantly for forced alignment of our conversational field recordings, the MAUS Italian system was developed through considerable training with spontaneous speech data. For more details, see Schiel et al. (2013).

3.1 Workflow for forced alignment Data processing is described here for (1) use of MAUS Italian system and (2) use of MAUS language independent system. Figure 1 shows the work flows for these two options. Starting from orthographic transcription and audio files, the procedures with thick solid arrows are shared by both options; those with thin solid arrows stand for MAUS Italian system; and those with dashed arrows stand for MAUS language independent system.

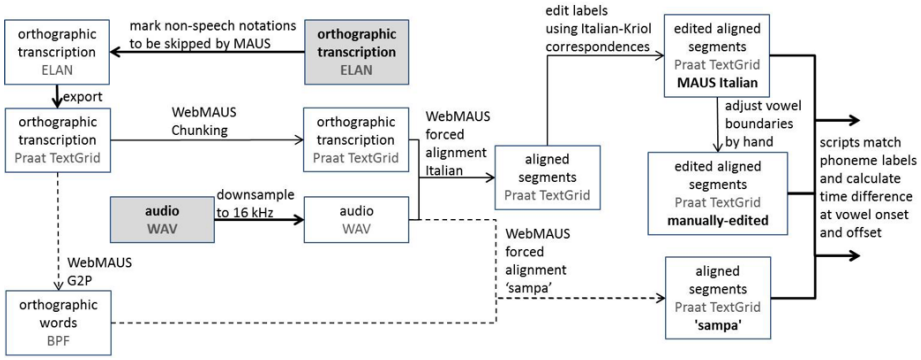

Figure 1. Sketch showing workflows for obtaining three datasets

3.1.1 MAUS Italian option In the five orthographic transcription files in ELAN, any non-speech notations in the transcription (e.g., notes about gestures, laughter or coughing, etc.) are marked with brackets to be skipped as *meta tag* by MAUS (WebMAUS General Help 2018). Following this, the transcription files were exported to Praat TextGrid files. The forced alignment process then involved three steps. First, the TextGrid files were run through the MAUS Chunking service, which offers better results for conversational data (Poerner & Schiel 2016). Second, the TextGrid and audio files (down-sampled to 16 kHz and chunked into segments of 20 MB maximum) were run through the WebMAUS forced alignment system for Italian (Schiel 1999; Kisler et al. 2016). Finally, the output TextGrid files containing the segment-level labels and alignment resulting from MAUS were downloaded locally.

In the five TextGrid files that were outputs of the forced alignment, the SAMPA segment labels (target vowel tokens and adjacent segments only) were edited using the Italian – Kriol correspondences in Table 1. Table 1 illustrates in the leftmost column the Italian SAMPA symbols which appeared when MAUS was run with the Kriol

Table 1

Italian (SAMPA)	Kriol (SAMPA)	Kriol (IPA)
i	I	ɪ
E	E	ɛ
e		
a	6	ɐ
O	O	ɔ
o		
u	U	ʊ
	{	æ
ei	eI	eɪ
ai	6I	ɐɪ
oi	oI	oɪ
au	6U	ɐʊ
	OU	oʊ
	i:	iː
	3:	ɜː
	u	uː
	6:	ɐː
	o:	oː
p	p	p
b	b	b
t	t	t
d	d	d
tS	tS	tʃ
ts		
dZ	dZ	dʒ
dz		
k	k	k
g	g	ɡ
f	f	f
	D	ð
s	s	s
z	z	z
S	S	ʃ
	h	h
m	m	m
n	n	n
J	J	ɲ
	N	ŋ
l	l	l
r	r\	ɹ
rr	r	r
w	w	w
j	j	j

orthographic input. (Not all Italian SAMPA symbols are shown since some were not produced as a result of the Kriol orthography which was input to the system). The editing of segment labels also involved *undoing* several phonological rules (e.g., word-initial velar voicing, germination) which had been applied by the WebMAUS Italian system (for details of these phonological rules see Schiel et al. 2013).

As can be seen in Table 1, there are consistent differences between the Italian SAMPA and the desired Kriol SAMPA. These made it feasible to run some simple find-and-replace substitutions to convert nearly all of the Italian SAMPA to Kriol SAMPA. The short vowels <i, e, a, o, u> were converted to <I, E, 6, O, U>. The diphthongs <ei, ai, oi, au> were converted to <eI, 6I, oI, 6U>. Four consonants were converted using context-free replacements: <r> to <r\>(version 6.0.23).[4]

3.1.2 MAUS language-independent option (*sampa*) In this processing, WebMAUS General was used for forced alignment with selection of the language-independent option (henceforth, *sampa*). The input to WebMAUS General was a BPF file created from a TextGrid (of the orthographic speaker tier exported from ELAN). The BPF file was created in two steps: first the G2P module of MAUS was used, to generate a template for the BPF file. This template contained (i) start and end times for each turn (TRN lines) with the utterance in orthography, (ii) a tokenised list of orthographic words in the order they appeared, one per line (ORT lines), and (iii) SAMPA versions of the tokenised list of orthographic words (KAN lines). The KAN lines were created using a Python script to edit the BPF file, which resulted from the use of G2P so that the KAN lines were close approximations to a phonemic SAMPA string for each word.

4. Alignment comparisons The accuracy results reported in §4 involve comparisons of the raw-alignment data from the MAUS options (Italian, sampa) with manually-edited alignment. The latter is intended as a type of "gold standard" human benchmark using familiar criteria for phonetic landmarks. The start and end timepoints for each target vowel interval were extracted from the manually-edited and raw TextGrid files using a Matlab script. This script first used the vowel label, the vowel start and end timepoints, and the sequence of adjacent phonemes to match each phoneme in the manually-edited TextGrid with its corresponding phoneme in the raw-alignment TextGrid. See Figure 2 for illustration.

A very small amount of the data was systematically excluded from analyses, as these tokens were uninterpretable. Eight tokens (n = 1050, 0.76%) in the Italian data do not match any corresponding labels in the manually-edited data and are thus excluded in the analysis in §4.1 below, giving $n_{Italian}$ = 1042. For reference, in relation to the results presented in §4.2, one token (0.10%) is excluded for sampa: n_{sampa} = 1049. With the matching information, the script calculated the time difference between each manually-edited vowel onset and the corresponding raw-alignment vowel onset, and likewise for the vowel offsets. The script also extracted the corresponding

[4]http://www.praat.org/.

orthographic word, utterance, and the preceding and following segments using the word and utterance alignment provided by WebMAUS.

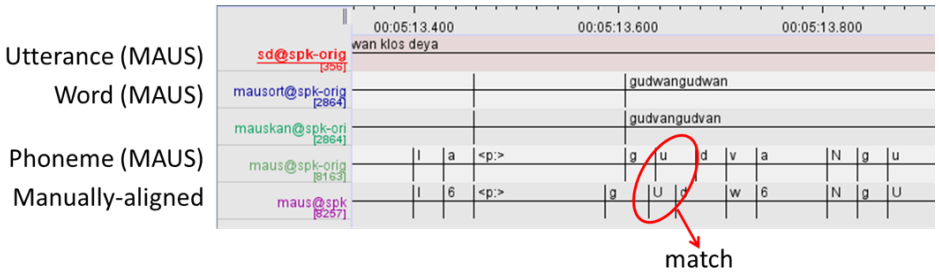

Figure 2. An example showing the match between manually-edited alignment and MAUS-Italian alignment

4.1 Comparing MAUS Italian to manually-edited In this section we report general alignment accuracy, comparing MAUS Italian to manually-edited (§4.1.1). We break down the results separately for the alignment of vowel boundaries at vowel onset vs. vowel offset. We would expect that vowel onset might be more accurately identified than vowel offset for the MAUS Italian system, since a stop burst is a relatively clear acoustic landmark in comparison to a fading vowel (and vowel onsets are more reliably identified than vowel offsets even by expert human aligners). We also explore in §4.1.2 the extent to which the alignments produced by MAUS Italian might have predictable errors as a function of other contextual factors, specifically vowel type (short vowel, long vowel, diphthong) and consonantal context (category of preceding or following consonant, in terms of consonant manner – stops, fricatives, nasals, liquids, and approximants). Again, we might expect predictably more inaccurate alignments for vowel boundaries when the following consonant is more vowel-like (more sonorous). Human aligners also face challenges in establishing vowel boundaries between vowels and liquids or approximants, for example. In §4.1.3 we report on the extent of edits to vowel labels that were required in the use of MAUS Italian for Barunga Kriol. This is a specific issue that arises in the use of a forced alignment system for a major world language (here, Italian) used "off the shelf" (i.e., in untrained form) for a low-resource language (here, Kriol).

4.1.1 General alignment accuracy Table 2 shows agreement at different thresholds (i.e., absolute time difference), separately for vowel onset and vowel offset as well as overall. The overall agreement of Italian with manually-edited alignments was 77.0% at 30 ms and 64.6% at 20 ms. These kinds of percentages indicate, for example, that 64.6% of the vowel boundaries in the Italian output fell within 20 ms of the placement of vowel boundaries by a human aligner (i.e., in the manually-edited version). These thresholds were chosen for comparison with the data reported in the forced alignment literature on "small" languages; 20 ms is the most stringent threshold applied in forced alignment literature, and the literature on "small" languages also

reports at the 30 (and 50) ms threshold. Agreement of Italian with manually-edited alignments was better at vowel onset than vowel offset (paired samples t[1041] = −3.8, p < 0.001; mean difference = −7.3 ms, with 95% confidence interval = −11.0, −3.5).

Table 2. Agreement of Italian and manually-edited alignments

Threshold	Vowel onset	Vowel offset	Overall
10 ms	45.8%	36.5%	41.2%
20 ms	72.1%	57.2%	64.6%
30 ms	82.3%	71.8%	77.0%
40 ms	86.3%	79.4%	82.8%
50 ms	88.7%	83.1%	85.9%

4.1.2 Effect of vowel type and segmental context Agreement at vowel onset and vowel offset was analysed with linear mixed effects models. The dependent variable in each model was the absolute difference (in ms) between the manually-edited boundary and the raw boundary (from MAUS Italian). We ran the models separately for vowel onset and vowel offset. For each factor of interest (vowel type, preceding context, following context), we compared a null model (i.e., intercept-only, using random intercept for speaker) with a model which included the relevant factor as the sole fixed effect, plus random slope for speaker. We used the anova function within the lme4 library to compare models, using the chi-square statistic, to assess the effect of each factor on alignment difference.

There are no significant effects of vowel type (short vowels, long vowels, diphthongs), using short vowels as the reference level, either at vowel onset ($\chi^2(7) = 4.9$, $p = 0.675$) or at vowel offset ($\chi^2(7) = 0.690$, $p = 0.998$). We inspected the alignment data using a 50 ms threshold,[5] to see which type of vowel or which type of context might tend to result in more misalignment. Long vowels tend to result in more misalignments, particularly at vowel onset, as shown in Figure 3. At vowel onset, 20.0% (37/185) of the boundaries of long vowels do not fall within 50 ms of the manually-edited boundary. The higher percentage of misalignments for long vowels may be due to the fact that the orthography for Barunga Kriol (and so the SAMPA labels) does not distinguish short from long vowels (as long vowels are rare and variable).

Among all the tokens with absolute time difference smaller than 50 ms, 513 tokens are between −50 ms and 0 ms and 411 between 0 ms and 50 ms for vowel onset, whereas those for vowel offset are 321 and 545 respectively. Figure 4 shows the time distribution of tokens whose time difference is larger than 50 ms (in the logarithmic form for ease of comparison). Similarly, there are more negative tokens for vowel onset, i.e. $t_{onset,Italian} - t_{onset,manually-edited} < 0$, and more positive tokens for vowel

[5]Here a 50 ms threshold was chosen for two reasons. First, for reliability analysis, the agreement rate is reported using 50 ms in the literature, e.g., DiCanio et al. 2013. Second, as shown in Figure 9, the agreement rate increased as the time difference increased but became more or less stable above 50 ms, indicating duration-independent misalignment that is worth further investigation.

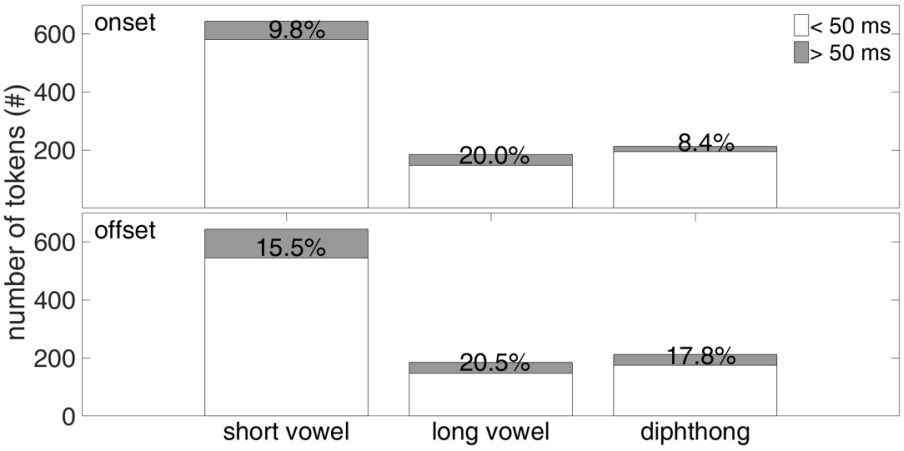

Figure 3. Number and proportion of boundaries accurate at 50 ms threshold, for Italian aligner by vowel type

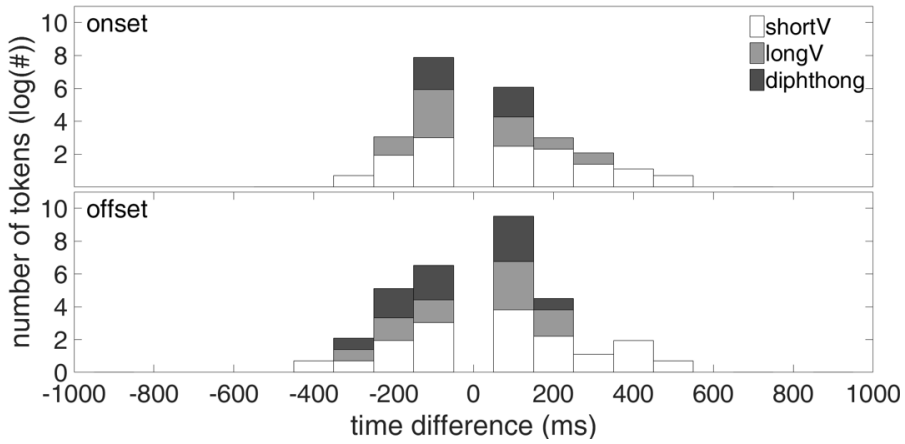

Figure 4. Distribution of tokens with time difference larger than 50 ms, grouped by vowel type

offset. MAUS has onset later and offset earlier than human, suggesting that different thresholds (different strategies) for locating boundaries are used by MAUS and humans.

For preceding and following context, agreement at vowel onset and vowel offset was analysed with linear mixed effects models, with the same approach described above. Neither vowel onset nor vowel offset alignment was related to type of preceding or following context (type of segment by manner, or silent interval). At vowel onset, there was no effect of preceding context ($\chi^2(42) = 18.3$, $p = 0.999$) or following context ($\chi^2(42) = 9.08$, $p = 1$). At vowel offset, there was no effect of preceding context ($\chi^2(42) = 14.1$, $p = 1$) or following context ($\chi^2(42) = 31.0$, $p = 0.895$). For these analyses we used approximant as the reference level.

Similar to Figure 3, Figure 5 shows the time difference between Italian and manually-edited by preceding and following context. The number of tokens in the affricate and vowel categories range from 1–7, so are not statistically meaningful. There is a tendency for larger disagreements in alignment when the preceding or following context is an approximant or a silent interval (but this difference does not reach significance, as just noted).

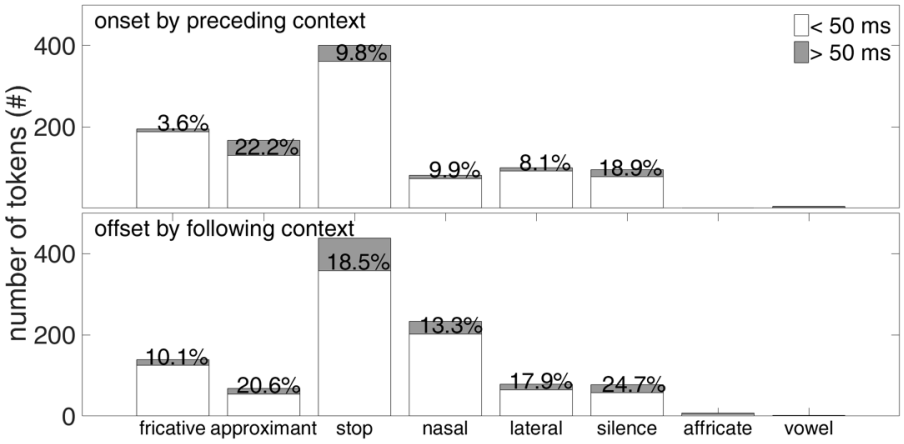

Figure 5. Time difference between Italian and manually-edited by preceding and following context

4.1.3 Edits to vowel labels Given the orthographic, phonological, and phonetic differences between Italian and Kriol, some phonemes were unavoidably labelled differently during the forced alignment using MAUS Italian. For the 1042 vowel tokens studied here, 454 of them had incorrect labels at the point of output from MAUS: 231 tokens (<{>, <OU>, and five long vowels, see Table 1) represent categories present only in Kriol; 215 tokens were labelled as <e> and <o>, which are unique in MAUS Italian. 26 diphthongs in Kriol were labelled as short vowels in Italian, 6 short vowels labelled as diphthongs, and 38 short vowels labelled as different short vowels. For the consonants, there is a total of 2982 tokens in the words from which the vowel tokens were extracted. 286 labels were found to be different from manually-edited ones: 188 tokens (65.7%) were matched to similar consonants, e.g., <g> in Kriol was labelled as <dZ>, <w> labelled as <v>, and <s> labelled as <z>; 34 tokens (11.9%) of <h> and <N> were mislabelled as they only exist in Kriol; and 35 tokens (12.2%) were missing in MAUS Italian.

4.2 Comparing sampa to manually-edited In this section we report the alignment accuracy and effects of vowel type and segmental context for the comparison of MAUS sampa with manually-edited. We report the same alignment comparisons as for Italian above (§4.1). There are no substantial vowel edits to report in the use of MAUS

sampa, since this option allows for specification of orthography-to-phoneme mappings which are relatively easy to define in a transparent orthography like Kriol.

4.2.1 General alignment accuracy Table 3 shows agreement at different thresholds (absolute time difference for vowel boundaries), separately for vowel onset and vowel offset as well as overall, for the comparison between sampa and manually-edited. Overall agreement was 65.6% at 30 ms and 53.7% at 20 ms.

Agreement of sampa with manually-edited alignments was better at vowel onset than vowel offset (paired samples t[1048] = −1.8, p = 0.0798; mean difference = −5.3 ms, with 95% confidence interval = −11.2, 0.6), similar to that of Italian with manually-edited. However, the overall agreement for sampa was about 10.7% lower than that for Italian at all thresholds, compared with the data in Table 2.

Table 3. Agreement of sampa and manually-edited alignments

Threshold	Vowel onset	Vowel offset	Overall
10 ms	37.9%	25.7%	31.8%
20 ms	60.5%	46.9%	53.7%
30 ms	69.8%	61.3%	65.6%
40 ms	73.4%	69.4%	71.4%
50 ms	77.0%	73.8%	75.4%

There are 241 and 275 tokens in sampa that are off by more than 50 ms at vowel onset and offset respectively, i.e., 104.2% and 56.2% more than those in Italian. Thus, based on manually-edited data, the sampa alignment is worse than Italian alignment at both vowel onset and offset, in terms of overall agreement at all thresholds from 10 ms to 50 ms and number of tokens with time difference larger than 50 ms. (We recognise, however, that the manually-edited alignment does derive from the Italian MAUS output, so a closer correspondence between manually-edited and Italian is to be expected, to a degree.)

4.2.2 Effect of vowel type and segmental context Mixed effects models (run as for the Italian alignment above) show no effect of vowel type at vowel onset ($\chi^2(7) = 8.0$, p = 0.331) or vowel offset ($\chi^2(7) = 4.4$, p = 0.729). Figure 6 shows the time difference at both vowel onset and offset between sampa and manually-edited by vowel type. Overall 31.4% of the long vowel alignments are off by more than 50 ms, contributing more to the outliers than short vowels and diphthongs. This could suggest that both MAUS Italian and MAUS sampa has least agreement with humans when determining the onset and offset of long vowels.

Similar to the results for MAUS Italian, there are more tokens with negative time difference at vowel onset and with positive time difference at vowel offset for sampa within the range of −50 ms to 50 ms: 525 versus 283 for 194 versus 580. Figure 7 shows the distribution of tokens with time difference larger than 50 ms at both vowel onset and vowel offset for sampa versus manually-edited data (in the logarithmic

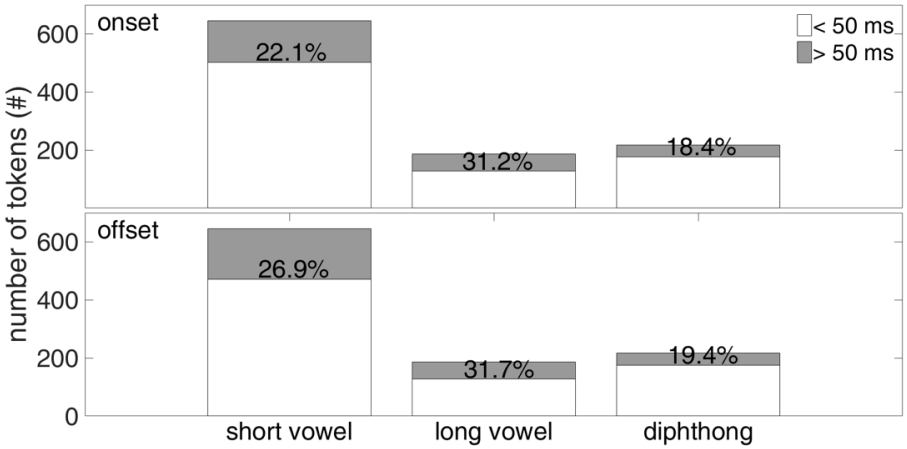

Figure 6. Time difference between sampa and manually-edited by vowel type

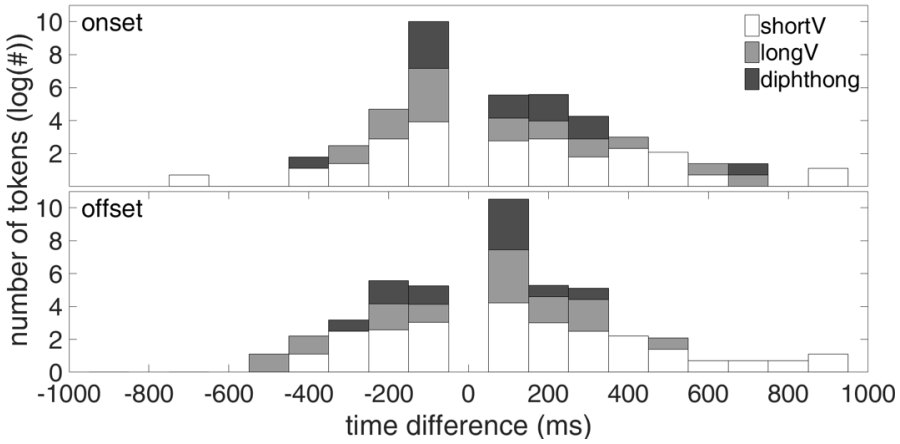

Figure 7. Distribution of tokens with time difference larger than 50 ms grouped by vowel type

form for ease of comparison). Qualitatively, MAUS sampa also puts onsets later and offsets earlier than human.

At vowel onset there is no effect of preceding context ($\chi^2(42) = 12.3$, $p = 1$), and no effect of following context ($\chi^2(42) = 30.2$, $p = 0.913$), using approximant as the reference level. At vowel offset there is no effect of preceding context ($\chi^2(42) = 9.1$, $p = 1$), nor of following context ($\chi^2(42) = 20.4$, $p = 0.998$), again using approximant as the reference level. Figure 8 shows the time difference between sampa and manually-edited by preceding and following context. The number on each bar is the percentage of tokens with absolute time difference larger than 50 ms.

For approximant, silence, and lateral contexts, sampa has least agreement. Compared with those shown in Figure 5, sampa has 6.3–21.8% more tokens distributed

with absolute time difference larger than 50 ms than Italian, regardless of the type of preceding and following context.

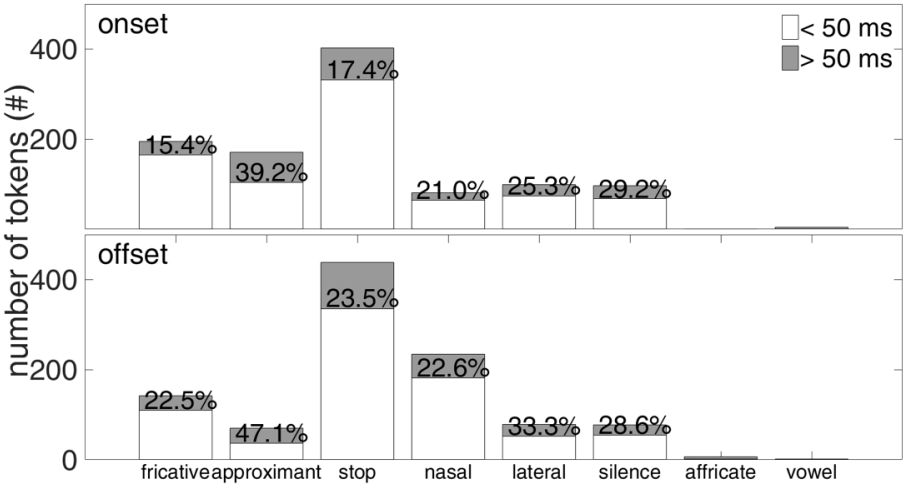

Figure 8. Time difference between sampa and manually-edited by preceding and following context

4.3 Data refinement by removing completely misaligned tokens As discussed above, neither MAUS Italian system nor MAUS language independent mode is originally designed for the forced alignment of north Australian Kriol. Unavoidably, there are missing, extra, and wrong phonetic labels (discussed in §4.1.3) and misaligned segments. In this study, the tokens with missing labels were excluded before further analysis. In some extreme cases, the onset and offset time can be off for a few seconds compared with the manually-edited data (which occurs for other automated aligners as well (MacKenzie & Turton 2013)). In our dataset we noticed that completely misaligned tokens tended to involve long stretches of sonorous segments (e.g., vowels, nasals, liquids, and glides) where presumably MAUS lacked strong acoustic landmarks like stop-vowel boundaries to assist in the alignment. We decided that it would be interesting to see how the time difference in vowel onset and offset varies with these cases excluded (following DiCanio et al. (2013), who excluded tokens with missing phones while comparing two machine aligners with human aligners in the analysis). In our data, 68 out of 1042 tokens (6.53%) for Italian and 151 out of 1049 (14.49%) for sampa were found to be completely misaligned, i.e., $t_{offset,MAUS} < t_{onset,manually-edited}$ or $t_{onset,MAUS} > t_{offset,manually-edited}$. Figure 9 shows the percentage of agreement with the misaligned data excluded for Italian and sampa respectively, compared with that of original data.

Once the completely misaligned data are removed, the time agreement between MAUS (for both Italian and sampa) and manually-edited alignment improves by a few percent at all time levels. For example, the agreement for Italian increases by 5.4% at 30 ms and 5.5% at 50 ms, while that for sampa increases by 11.0% and 12.4% re-

spectively. See Table 4 for key comparisons between Italian and sampa (overall, data at vowel onsets and offsets combined). These results suggest that the time difference at vowel onset and offset between MAUS alignment and manually-edited alignment is partly due to the complete misalignment at phonetic level, especially for sampa.

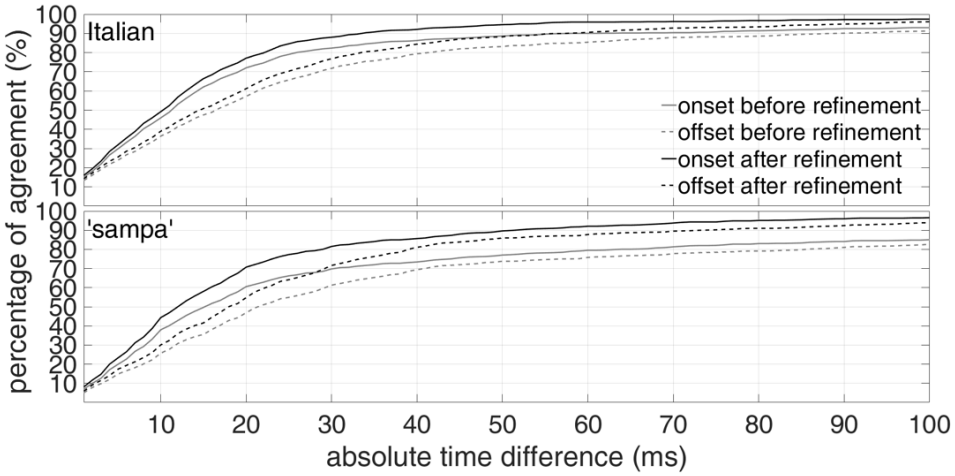

Figure 9. Agreement at different thresholds improved with completely misaligned tokens excluded

Table 4. Agreement results after data refinement

Forced alignment option	Threshold	Agreement overall	Agreement improvement
Italian	20 ms	69.2%	4.6%
	30 ms	82.4%	5.4%
	50 ms	91.4%	5.5%
'sampa'	20 ms	62.8%	9.1%
	30 ms	76.6%	11.0%
	50 ms	87.8%	12.4%

5. Discussion and Conclusions The goal of the present study was to provide an updated and detailed evaluation of forced alignment on data in north Australian Kriol, a "small" language, without training on that language. Previous research using this kind of approach, dubbed *cross-linguistic forced alignment* (CLFA; Kempton et al. 2011; Kempton 2017) or *untrained* alignment (DiCanio et al. 2013), had reported encouraging results, with disagreement rates of 34–51% at 20 ms. The higher disagreement rates have typically been associated with the alignment of conversational speech (Kurtic et al. 2012) and/or speech that has background noise or transcription that is incomplete or relatively less accurate (Strunk et al. 2014).

The results of using MAUS system trained on Italian, a major world language whose MAUS training dataset included conversational recordings, were at least as good as and possibly better than previous results using CLFA.[6] In previous evaluation with conversational data, error rates were >50% at 20 ms. In our data, at vowel onset the error at 20 ms was 27.9% and at vowel offset was 42.8%. These figures are conservative, as they do not reflect corrections applied by removing outliers (as done in the analysis by DiCanio et al. (2013)). When such corrections are made, the disagreement rates at 20 ms are less than 22.9% and are substantially lower still at 30 ms, probably because the completely misaligned tokens are less likely to fall within the 20 ms threshold. At 50 ms, the disagreement rate is under 10%. These are highly adequate results for a field linguist interested in vowel analysis, for instance, who needs to correct relatively few vowel boundaries before running spectral and temporal analyses of vowels. An obvious disadvantage of this approach, however, is the need to edit vowel labels to reflect the spelling of the actual language rather than the MAUS major world language. For a language like Kriol, where there is no completely regular spelling system and varieties differ in how they are spelled, this is less concerning, as even a system that was trained for Kriol would require manual checks of the labels. This would presumably be true for many other under-resourced languages where the language is less well understood, and the phonological analysis and/or orthography is basic to non-existent.

The other clear option is to use MAUS in language-independent mode. This ensures more accurate phone labels are generated from the orthography (compared with 43.6% labelling error in MAUS Italian). How good is the alignment? Our results show higher disagreement rates for the MAUS sampa mode than for Italian mode: at vowel onset the error rate at 20 ms is 39.5% (rather than 27.9%) and at vowel offset it is 53.1% (rather than 42.8%). These are not perhaps enormously different, though in practical terms they are; in a corpus containing 1,000 vowel tokens for analysis the number of additional boundaries requiring manual editing would be 219, for example. There are also more major misalignments with the sampa mode than the Italian mode, i.e., alignments that are very far away from the "gold standard" (> 50 ms, up to several seconds away). These are concerning because they tend to take even longer to manually edit the alignment.

It is to be hoped that it will become easier for regular linguists and/or community members working on under-resourced languages to train language-specific models for forced alignment, through easier interfaces and reduced needs for large training datasets. The report and recommendations recently offered for the use of ProsodyCat (Johnson et al. 2018) is a good example of this promise. Another option that holds promise is investment in the training of acoustic models for language families, or groups of languages which are phonologically and orthographically similar, such as Australian languages (Stoakes & Schiel 2017). Meanwhile, our research has updated

[6]One caveat to our results is that we have, like previous research in CLFA, used categorical boundary threshold data rather than continuous measurement of segment overlap that is duration-independent. For more details on the latter analytical option see Figure 2 in Paulo & Oliveira (2004).

the evaluation of cross-linguistic forced alignment for an under-resourced language, with results that appear to be more accurate than some earlier results.

References

Adams, Oliver. 2018. *Automatic understanding of unwritten languages*. Melbourne: The University of Melbourne. (Ph.D. thesis).

DiCanio, Christian, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, & Rey 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. doi:10.1121/1.4816491.

Johnson, Lisa M., Marianna Di Paolo, & Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation & Conservation* 12(1). 80–123. http://hdl.handle.net/10125/24763.

Jones, Caroline, Katherine Demuth, Weicong Li, & André Almeida. 2017. Vowels in the Barunga variety of North Australian Kriol. *Interspeech 2017*. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, August 20–24, 2017. 219–223. doi:10.21437/Interspeech.2017-1552.

Kempton, Timothy. 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. Paper presented at the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2), Honolulu, Hawaii, March 6–7, 2017.

Kempton, Timothy, Roger K. Moore, & Thomas Hain. 2011. Cross-language phone recognition when the target language phoneme inventory is not known. *Interspeech 2011. In Proceedings of the 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, August 28–31, 2011. 3165–3168.

Kisler, Thomas, Uwe D. Reichel, & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. doi:10.1016/j.csl.2017.01.005.

Kisler, Thomas, Uwe D. Reichel, Florian Schiel, Christoph Draxler, & Bernhard Jackl. 2016. BAS speech science web services – an update of current developments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May 23–28, 2016. 3880–3885. https://www.phonetik.uni-muenchen.de/ reichelu/publications/KRSDJP_LREC2016.pdf.

Kurtic, Emina, Bill Wells, Guy J. Brown, Timothy Kempton, & Ahmet Aker. 2012. A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, Istanbul, Turkey, May 23–25, 2012. 1323–1327. http://www.lrec-conf.org/proceedings/lrec2012/pdf/513_Paper.pdf.

MacKenzie, Laurel & Danielle Turton. 2013. Crossing the pond: Extending automatic alignment techniques to British English dialect data. Presented at the 42nd annual meeting of the New Ways of Analyzing Variation (NWAV), Manchester, England, October 20, 2013.

Paulo, Sérgio & Luís C. Oliveira. 2004. Automatic phonetic alignment and its confidence measures. In *Proceedings of the 4th International Conference on Natural Language Processing: Advances in Natural Language Processing*, EsTAL 2004, Alicante, Spain, October 20–22, 2004. 36–44 doi:10.1007/978-3-540-30228-5_4.

Poerner, Nina & Florian Schiel. 2016. An automatic chunk segmentation tool for long transcribed speech recordings. 12. *Tagung Phonetik und Phonologie im deutschsprachigen Raum*, Munich, Germany, October 12–14, 2016. 145–147.

Schiel, Florian. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, CA, August 1–7, 1999. 607–610. https://epub.ub.uni-muenchen.de/13682/1/schiel_13682.pdf.

Schiel, Florian, Susanne Burger, Anja Geumann, & Karl Weilhammer. 1997. The Partitur format at BAS. *Forschungsberichte-Institut für Phonetik und Sprachliche Kommunikation der Universität München* 35. 127–137. http://www.phonetik.uni-muenchen.de/forschung/FIPKM/vol35/f35_fs_2.pdf.

Schiel, Florian, Mary Stevens, Uwe D. Reichel, & Francesco Cutugno. 2013. Machine learning of probabilistic phonological pronunciation rules from the Italian CLIPS Corpus. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, August 25–29, 2013. 1414–1418. https://epub.ub.uni-muenchen.de/18046/1/SchielIS2013.pdf.

Stoakes, Hywel & Florian Schiel. 2017. A Pan-Australian acoustic model: Automatic alignment using the MAUS. Conference of the Australian Linguistic Society 2017, Sydney, Australia, December 4–7, 2017.

Strunk, Jan, Florian Schiel, & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, May 26–31, 2014. 3940–3947. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1176_Paper.pdf.

WebMAUS General Help. 2018. https://clarin.phonetik.uni-muenchen.de/BASWebServices/#!/help. (Retrieved 12 March 2018).

Caroline Jones
caroline.jones@westernsydney.edu.au
orcid.org/0000-0001-6277-8262

Weicong Li
Weicong.Li@westernsydney.edu.au
orcid.org/0000-0002-7423-2846

Andre Almeida
a.almeida@unsw.edu.au
orcid.org/0000-0001-6075-7281

Amit German
amitjgerman@gmail.com