Nor Jnl Ling 00.0, 1–27 © Nordic Association of Linguists doi: 00000000000000

ARTICLE

Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype

Nathan J. Young & Michael McGarrah

We propose a rapid adaptation of FAVE-Align to the Nordic languages, and we offer our own adaptation to Swedish as a template. This study is motivated by the fact that researchers of lesser-studied languages often neither have sufficient speech material nor sufficient time to train a forced aligner. Faced with a similar problem, we made a limited number of surface changes to FAVE-Align so that it – along with its original hidden Markov models for English – could be used on Stockholm Swedish. We tested the performance of this prototype on the three main sociolects of Stockholm Swedish and found that read-aloud alignments met all of the minimal benchmarks set by the literature. Spontaneous-speech alignments met three of the four minimal benchmarks. We conclude that an adaptation such as ours would especially suit laboratory experiments in Nordic phonetics that rely on elicited speech.

Keywords computational automatic speech recognition tools, forced alignment, sociolinguistics, sociophonetics, Nordic dialectology, Swedish varieties

Nathan J. Young, Centre for Research on Bilingualism, Stockholm University, Stockholm 106 91, Sweden. nathan.young@biling.su.se
Michael McGarrah, Department of Computer Science, Georgia Institute of Technology, North Avenue, Atlanta, Georgia 30332, USA. mcgarrah@gmail.com

1. Introduction

When conducting phonetic investigations of a lesser-studied language, researchers will often encounter resource challenges when it comes to segmentation. Even for Swedish, a language not typically considered lesser-studied, very few technological tools circulate for phoneticians. This paper seeks to address this gap by incorporating a simple and straight-forward Swedish-language adaptation of FAVE-Align (Rosenfelder et al., 2011) that resembles approaches used in the past for endangered languages (DiCanio et al., 2013; Coto-Solano and Solórzano, 2017; Coto-Solano et al., 2018; Strunk et al., 2014). The novel contributions of this paper are that (1) this is the first published adaptation for any Nordic language, (2) this adaptation meets most of the accuracy benchmarks established by the literature, (3) a step-by-step guideline is offered in the Appendix for those who wish to duplicate the adaptation for another Nordic language.

In cases where a language has not yet been modeled for forced alignment – or it has been modeled but not disseminated publicly[1] – phoneticians must invest in training a new aligner. Not only does this demand time and expertise, researchers

may not have access to sufficient transcribed material in that language, which is a key prerequisite for model training. But even if this material were to be available, such a task is always a potential 'rabbit hole' if the end product does not end up being sufficiently accurate. In other words, vetting the software *before* investing time in learning, training and validating is simply not possible, because a good track record for one language does not guarantee a similar track record for another (see, e.g., the various languages in Strunk et al., 2014). In the case of researchers and students working with small datasets where the material is insufficiently large for training, the frustrating reality is that manual alignment is often the only option.

This paper proposes an alternative; namely, adapting the English-language FAVE-Align to the Nordic languages while using its existing hidden Markov models. Whereas using such 'untrained' models has rendered unreliable results for endangered languages typologically distant from the original language(s) used for training (DiCanio et al., 2013; Coto-Solano and Solórzano, 2017; Coto-Solano et al., 2018; Strunk et al., 2014), we show it to be robust and reliable for spontaneous and read-aloud Stockholm Swedish – likely because the variety is more typologically similar to English. In crudely and quickly adapting FAVE-Align to Stockholm Swedish, we were able to reduce total manual segmentation time to approximately 78 hours per recorded hour. For spontaneous speech, 37 percent of the boundaries fell within 10 milliseconds and 65 percent of the boundaries fell within 20 milliseconds of the manual-alignment benchmark. For read-aloud speech, these figures were 50 percent and 73 percent, respectively. Successful alignment of spontaneous speech requires of course access to a comprehensive pronunciation dictionary, and this is not always available for lesser-studied languages. However, aligning read-aloud speech requires just a short list of pronunciations, so we believe that our latter results will have the widest reach.

Given the long absence of publicly-disseminated forced aligners for any of the Nordic Languages[2] and the fact that untested Swedish aligners have only recently been released (see Section 2.2 for a review), this paper can serve both as a methodological template for adapting FAVE-Align to other Nordic varieties (see Appendix) and as a base reference for benchmarking the performance of future aligners. Such peer-reviewed benchmarks are needed as linguists pollinate technological movement in the field, and they are vital for seeking out prospective grants and funds to finance the training of designated Nordic-language aligners.

2. Background

2.1 Forced alignment and its advancement of phonetic research

With the help of readily-available forced-alignment programs, phonetic investigations of English have advanced further than those of any other language. Meanwhile, phonetic investigations of the Nordic languages, including Swedish, have lagged. To offer an example, we examined and coded – according to language researched – the 782 articles published between 2001 and 2020 in the Journal of Phonetics. The top three researched languages were English, German and French with 336, 71, and

65 articles, respectively. Swedish, the most-commonly researched Nordic language, had a mere 12 articles, followed by Norwegian with eight, Danish with four, and Icelandic with one. Proportionate to number of speakers, these languages are somewhat underrepresented. Finnish, a language with approximately 5 million speakers, had 19 articles; Arrernte, a language with approximately 4 000 speakers, had six articles. As an additional example, before the onset of the project to which this development is tied (Young 2019), only three variationist investigations had ever been conducted on Swedish (Gross et al., 2016; Kotsinas, 1994; Nordberg, 1975). Among these three, only the first-listed study was acoustic-phonetic, relying on manual segmentation (personal conversation with Johan Gross, 2020). The latter two were based on data that was never phonetically segmented; rather, variants were perceptually coded and counted.

We believe a circular dynamic is at play. The low number of contemporary phoneticians engaged with research on Nordic languages[3] has translated into few investments in forced alignment. In turn, this lack of investment has perhaps discouraged growth of the field. For English, the same feedback cycle may also be operating, albeit in the opposite direction. The early dominance of research on English has motivated the development of a high number of forced aligners, which has allowed the anglo-linguistic enterprise to be more prolific than ever.

The four main forced-alignment suites that circulate today were all trained on the English language. They are *Forced Alignment and Vowel Extraction* (*FAVE-Align*, Rosenfelder et al. 2011; Yuan et al. 2013), *ProsodyLab aligner* (Gorman et al., 2011), *LaBB-CAT Transcriber* (Fromont and Hay, 2012), and the *Montreal Forced Aligner* (McAuliffe et al., 2017). FAVE-Align (formerly known as the Penn Forced Aligner, Yuan and Liberman 2008), ProsodyLab aligner, and the Montreal Forced Aligner are modeled on American English. LaBB-CAT is modeled on New Zealand English.

As a very recent addition (and after the onset of the present study), the Montreal Forced Aligner began offering pre-trained acoustic models for Bulgarian, Croatian, Czech, French, German, Hausa, Korean, Mandarin, Polish, Portuguese, Russian, Spanish, Swahili, Swedish, Thai, Turkish, Ukrainian, and Vietnamese. These newer models are trained on read-aloud speech and require the use of the GlobalPhone dictionary (Schultz and Schlippe, 2014), which is proprietary and costs 600 euros to obtain (alternative pronunciation dictionaries cannot be used because phone coding within the models is opaque). Performance metrics have not yet been released for any of these newer models (see Section 2.4).

Other options are *EasyAlign* for Praat (Goldman, 2011) and the *BAS Speech Science Web Services* (Kisler et al., 2016). EasyAlign offers automatic transcription for French, Spanish, and Taiwan Min, and works only on Windows machines. It appears that a singular adaptation had been made for Swedish in 2007, but this adaptation has not been made available to the public, and performance metrics were not ever disclosed (Lindh, 2007). BAS Speech Science Web Services has offered for quite some time a web-accessible interface called *WebMAUS Basic* for automatic transcription of Basque, Catalan, Dutch, English, Estonian, Finnish, Georgian, German, Japanese, Hungarian, Italian, Maltese, Polish, Russian, and Spanish. Recently and

also after the onset of this project, Swedish was also added, but performance metrics have not been released on this either.

As has been discussed in the Introduction, the present study is not the first time that FAVE-Align has been adapted as an 'untrained' prototype for lesser-studied languages. DiCanio et al. (2013) built an adaptation for Yoloxóchitl Mixtec, and Coto-Solano and Solórzano (2017) built a similar prototype for the endangered language Bribri and, later, Cook Islands Maori (Coto-Solano et al., 2018). Strunk et al. (2014) built a language-general model and used it to align read-aloud and spontaneous Baura, Bora, Even, and Sri Lankan Malay. These aligners produced between poor and fair accuracy levels, likely due to the typological difference between them and the language(s) their respective aligners were trained on (mostly Indo-European). The present adaptation stands out from this group in that its accuracy performance is competitive with custom-trained aligners.

What this review aims to demonstrate is that the development of forced alignment programs has been solidly anglocentric and that the expansion to other languages has excluded Nordic languages until very recently. This curious exclusion motivated the present study and has likely motivated the recent addition of Swedish to MFA and WebMAUS Basic. The ensuing performance analysis will serve as a handy baseline benchmark for the eventual testing of these other Swedish adaptations, and the step-by-step instructions we provide will allow others to duplicate our adaptation for other Nordic languages.

2.2 How forced alignment works

ProsodyLab, FAVE-Align, LaBB-CAT, BAS, and EasyAlign rely on the proprietary Hidden Markov toolkit (a.k.a. HTK; Young et al. 1993), and the Montreal Forced Aligner (MFA) relies on the open-source Kaldi (Povey et al., 2011), which is a type of neural network. Regardless of program, the inputs are always (1) an orthographic transcription, (2) a sound file, and (3) a pronunciation dictionary. The output is a phonetically-segmented file for use in Praat (Boersma and Weenink, 2017). The orthographic transcription is often a tab-delimited file outputted by ELAN (Sloetjes and Wittenburg, 2008) that has start and end times for each phrase/breath group (see Figure 1). The pronunciation dictionary is a text file that has pronunciation entries for every word in the language, which often can be as high as 30 or 40 possibilities for long compound words. This can be seen in Figure 2 where '*cirkusartist*' has a canonical pronunciation option like [ˈsɪr.kʉs.aˌtɪst] on line 21, but a series of elided options such as [ˈsɪr.kʉ̟.ˌtɪs] on line 14[4] . The final output, exemplified in Figure 3, is a Textgrid file for use with Boersma and Weenink's (2017) Praat that, as we discuss in the following sections, can vary in accuracy depending on the aligner at hand.

Most of the programs are free of cost (with the exception of the GlobalPhone extension of the MFA), and they provide various amounts of source code to the public along with varying degrees of written instructions for customizing the software to new languages. FAVE-Align stands out because it was specifically designed for sociolinguistic purposes and because it has shown the highest accuracy rates for the

```
August   0.0    2.373  Cirkusen var på väg! Deras plakat,
August   2.373  4.401  med bilden av en flygande cirkusartist
August   4.401  6.114  var uppsatta över hela stan.
August   6.114  9.426  Tidigare på dagen satt jag uttråkad när morfar
      ↪ ringde å  sa:
August   9.426  11.499 Jag har en överraskning till dig
August   11.499 13.862 något bra för humöret åsjälen
August   13.862 15.611 Vill du veta vad de är?
August   15.611 19.015 Ja! Berätta, berätta, berätta!
August   19.015 20.462 ropade jag förtjust.
August   20.462 22.302 Du får se själv i kväll.
August   22.302 23.797 Vad kul!
```

Figure 1. INPUTS 1 and 2: Five-column tab-delimited transcription input for FAVE-Align, produced with ELAN, and sound file.

```
1  AV AA1 V
2  BILDEN B IH1 L D EHO
3  BILDEN B IH1 L D EHO N
4  CIRKUSARTIST S IH3 K RS RT IH2 S
5  CIRKUSARTIST S IH3 K RS RT IH2 S T
6  CIRKUSARTIST S IH3 K S RT IH2 S
7  CIRKUSARTIST S IH3 K S RT IH2 S T
8  CIRKUSARTIST S IH3 K UHO RS RT IH2 S
9  CIRKUSARTIST S IH3 K UHO RS RT IH2 S T
10 CIRKUSARTIST S IH3 K UHO S AHO RT IH2 S
11 CIRKUSARTIST S IH3 K UHO S AHO RT IH2 S T
12 CIRKUSARTIST S IH3 K UHO S RT IH2 S
13 CIRKUSARTIST S IH3 K UHO S RT IH2 S T
14 CIRKUSARTIST S IH3 R K RS RT IH2 S
15 CIRKUSARTIST S IH3 R K RS RT IH2 S T
16 CIRKUSARTIST S IH3 R K S RT IH2 S
17 CIRKUSARTIST S IH3 R K S RT IH2 S T
18 CIRKUSARTIST S IH3 R K UHO RS RT IH2 S
19 CIRKUSARTIST S IH3 R K UHO RS RT IH2 S T
20 CIRKUSARTIST S IH3 R K UHO S AHO RT IH2 S
21 CIRKUSARTIST S IH3 R K UHO S AHO RT IH2 S T
22 CIRKUSARTIST S IH3 R K UHO S RT IH2 S
23 CIRKUSARTIST S IH3 R K UHO S RT IH2 S T
24 CIRKUSEN S IH1 K UHO S EHO
25 CIRKUSEN S IH1 K UHO S EHO N
26 CIRKUSEN S IH1 R K UHO S EHO
27 CIRKUSEN S IH1 R K UHO S EHO N
28 DERAS D EE3 R AHO S
29 EN AEH1 N
30 EN EE1 N
31 FLYGANDE F L YY3 G AHO N EHO
32 FLYGANDE F L YY3 G AHO D EHO
33 FLYGANDE F L YY3 G AHO N D EHO
34 MED M AEH1
35 MED M EE1 D
36 PØ_V$G P OAHO V AE1 G
37 PØ_V$G P V AE1 G
38 PLAKAT P L AHO K AA1 T
39 VAR V AA1
40 VAR V AA1 R
```

Figure 2. INPUT 3: Pronunciation dictionary with all possible pronunciations using ASCII characters for IPA.

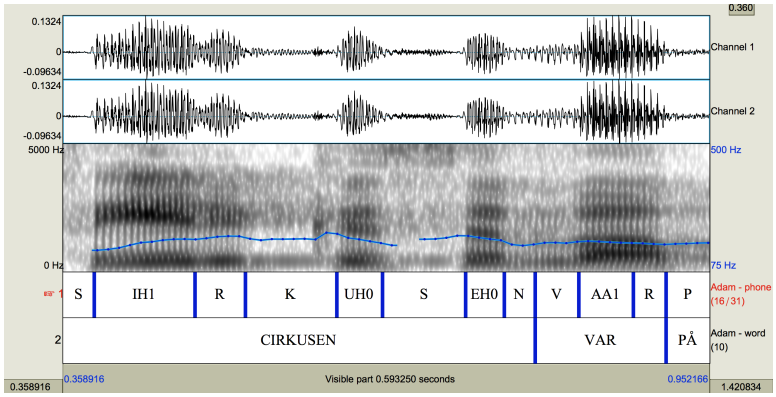

Figure 3. OUTPUT: Phonetically segmented file that is readable in Praat.

alignment of spontaneous vernacular speech (Yuan et al., 2013). Prosodylab and the MFA stand out because they provide the most robust assistance for training new languages. Additionally, the MFA is wrapped, which means it can be used out of the box with no subsidiary installations (e.g., Python).

FAVE-Align and the MFA are also noteworthy because they can process large sound files. They break files into chunks, align them, and concatenate them back together – all behind the scenes. This is very useful for any large-scale sociolinguistic project, but obviously less important for small projects. The remaining other programs require the user to manually break sound and transcription files down into one file per breath group. MFA is additionally noteworthy because it is relatively user-friendly and has an out-of-the-box trainer for new languages (should one have sufficient transcribed material handy).

We have offered a review of the various aligners on the market because this paper is, after all, about forced alignment. We would like, however, to point out that in the case of the Nordic languages, the comparative merits of each aligner do not matter much. In the case of Swedish, we had neither sufficient material to train an aligner like the MFA, and there were no pre-trained models available (and even today, the MFA model for Swedish sits behind a paywall). The picture is the same for Danish, Estonian Swedish, Faroese, Fenno-Swedish, Icelandic, the Northern and Western Norwegian dialects, and Övdalian. Absent of a large corpus of transcribed material, researchers will not be able to use MFA's out-of-the-box trainer. The only reasonable alternative is the one we propose here.

2.3 Teasing apart the benefits of forced alignment

The purpose of this paper is to share a resource – a FAVE 'hack', if you will – to help phoneticians save time. Therefore, we will first devote this section to unpacking where exactly the most time is spent in the segmentation process. In doing so, we hope to demonstrate convincingly that there is a limit to the additional amount of time one can save after a certain accuracy threshold.

There is indeed a consistent positive relationship between alignment accuracy and time saved – if one wishes to extract data from uncorrected files, which is often the practice for variationist projects that take formant measurements from the nucleus of, for example, 25 000-plus vowels within a corpus (Dodsworth and Benton, 2017, 377). However, for analyses of rhythm (Torgersen and Szakay, 2012; Thomas and Carter, 2006; Young, 2019), manual corrections are obligatory. Laboratory Phonetics studies, typically using smaller datasets, also mandate manually-aligned datasets (Chodroff and Wilson 2017, 33; Cole et al. 2019, 120). In such instances, the time needed to manually move an incorrect boundary is roughly the same for 5 milliseconds off-mark as it is for 40 milliseconds off-mark. What saves time is *fewer* inaccurate boundary placements, with the degree of accuracy being more or less unimportant once the boundary error crosses a pre-established threshold.

Importantly, those time savings are marginal when compared to the time needed to manually build boundaries and populate the resulting cells with the appropriate

phonetic orthography. To illustrate what we mean, take the following example. The recording that contains the first breath group from Figure 1 '*cirkusen var på väg. Deras plakat*'[5] lasts 2.37 seconds. We set a timer while the first author conducted the following tasks in Praat:

1. Building tiers then boundaries between words; populating the resulting cells: **2m 26s**
2. Building boundaries between phonemes; populating the resulting cells: **4m 24s**
3. Proofing boundaries; making final edits: **2m 19s**

It takes 9 minutes 9 seconds (**549 seconds**) to manually align a 2.37-second transcription, which makes our segmentation-to-recording ratio ratio 232:1[6] . Observe, however, that more than 75 percent of that time is spent building the boundary architecture and populating cells. Any program that can automatically do that has the potential to save a lot of time, regardless of how accurate boundary placement is. A program that can accurately place the boundaries is also a boon, but that is in many respects a secondary benefit.

It is this fact that motivated our choice to build a prototype from FAVE-Align rather than training an entirely new model for Swedish. Since we had no guarantee for future alignment accuracy, we felt that the rapid adaptation of a pre-existing aligner was the more prudent investment to make, since it would eliminate steps 1 and 2 no matter what. This is also the viewpoint taken by the researchers who paved the way for this study and used untrained aligners for typologically-rare endangered languages (DiCanio et al., 2013; Coto-Solano and Solórzano, 2017; Coto-Solano et al., 2018; Strunk et al., 2014). Although the accuracy levels were poor, they saved the authors considerable time in their alignment endeavors, nonetheless.

2.4 Identifying acceptable accuracy benchmarks

If one accepts the review presented in the aforementioned section, then nearly any level of accuracy is acceptable as a starting point from which to manually correct boundaries. Of course, the literature on forced alignment is not as permissive. It has established a consistent range of performance metrics that are reviewed below. We will later apply these same metrics as a way to assess the quality of our Swedish-language adaptation.

Many metrics circulate, and this can often make cross-comparability within the literature challenging. This paper will therefore limit itself to the four most-common metrics: (1) median onset difference from manual alignments, (2) mean onset difference from manual alignments, (3) the percentage of boundaries that fall within 10 milliseconds of the manual alignment, and (4) the percentage of boundaries that fall within 20 milliseconds of the manual alignment.

As it pertains to medians and means, some studies have solely calculated them for vowels (Evanini 2009, 56) or have calculated them from log-transformed absolute values (Wilbanks, 2015; Gorman et al., 2011). Here, we calculate them for all phonemes. Some studies have also used standard deviations (Labov et al., 2013) or the percentage of boundaries that fall within 5, 25, 30, 40, 50, and 100 mil-

liseconds of the manual benchmark (Cosi et al. 1991, 695; McAuliffe et al. 2017, 500). We chose not to include these metrics because their adoption is not sufficiently widespread. The below review will first cover automatic alignment benchmarks followed by manual alignment benchmarks. In select cases where visual figures are provided with no actual number, we have estimated the number by lining a straight edge between the plot and the axes (Evanini, 2009; Yuan and Liberman, 2008; Cosi et al., 1991). While different papers have rounded to different decimal levels, we round to the nearest whole percentage or millisecond.

2.4.1 Benchmarks for automatic alignment

Table 1 contains a schedule of the benchmarks laid out in the literature for forced alignment that we will discuss in the ensuing prose.

Cosi et al. (1991) is the earliest paper on phonetic forced alignment that we are aware of. They built an aligner for spontaneous Italian speech that had a mean error of 27 milliseconds when compared to manually-aligned boundaries. For the 10-millisecond and 20-millisecond benchmarks, they were able to achieve circa 41 percent for the former and between 57 percent and 64 percent for the latter (1991, 695).

Yuan and Liberman (2008), the most commonly-cited study for FAVE-Align, reported that approximately 70 percent of the boundaries fell within 10 milliseconds of the manual standard and that approximately 80 percent of the boundaries fell within 20 milliseconds of the manual standard (2008, 4). These measurements were calculated on the original US Supreme Court Justice corpus upon which FAVE-Align was also modeled. In later work, Yuan et al. (2013, 2308) proposed explicit phone boundary models within the Hidden Markov Model framework that improved the accuracy to 78 percent and 94 percent for the 10 and 20-millisecond error ranges, respectively.

Gorman et al. (2011) compared the performance of FAVE-Align with their newly-developed ProsodyLab Aligner on spontaneous American English taken from a television media corpus. They found FAVE-Align to have a median boundary error of 12 milliseconds and a mean boundary error of 21 milliseconds. For ProsodyLab, this was 12 and 31, respectively. Ten and 20-millisecond benchmarks were not calculated.

McAuliffe et al. (2017) assessed FAVE-Align and their newly-proposed Montreal Forced Aligner on read-aloud and spontaneous American English. For spontaneous speech run through FAVE-Align, the mean error was 19 milliseconds, and the median error was 12 milliseconds (2017, 501). For read-aloud speech run through FAVE-Align, the mean error was 22 milliseconds, and the median error was 13 milliseconds. Boundary-threshold percentages were not reported for FAVE-Align; they were, however, reported for the Montreal Forced Aligner. These were 41 percent within 10 milliseconds for spontaneous speech and 36 percent within 10 milliseconds for read-aloud speech (2017, 500). Twenty-millisecond thresholds were not calculated. What is particularly interesting about these results is that read-aloud speech aligned *less* accurately than spontaneous speech.

reference	tool	language	speech style	median displacement (ms)	mean displacement (ms)	pct within 10 ms	pct within 20 ms
— FORCED ALIGNMENT —							
Cosi et al. 1991	unnamed	Italian	S	-	27	41	57–64
Yuan & Liberman 2008	FAVE	AE	S	-	-	70	80
Yuan et al. 2013	FAVE	AE	S	-	-	78	94
Gorman et al. 2011	FAVE	AE	S	12	21	-	-
Gorman et al. 2011	ProsodyLab	AE	S	12	31	-	-
McAuliffe et al. 2017	FAVE	AE	S	12	19	-	-
McAuliffe et al. 2017	MFA	AE	S	-	-	41	-
MacKenzie & Turton 2020	FAVE	BE	S	-	8–20	-	76–90
Goldman 2011	EasyAlign	AE	S	-	-	50, 51	75, 77
Goldman 2011	EasyAlign	French	S	-	-	49, 52	79, 82
Wilbanks 2015	FASE	Spanish	S	-	21	45	70
McAuliffe et al. 2017	FAVE	AE	R	13	22	-	-
McAuliffe et al. 2017	MFA	AE	R	-	-	36	-
MacKenzie & Turton 2020	FAVE	BE	R	-	8	-	83
Hosom 2009	unnamed	AE	R	-	-	80	93
	Lower bound in literature			13	31	36	57
	Upper bound in literature			12	8	80	94
— MANUAL ALIGNMENT —							
Cosi et al. 1991	manual	Italian	S	-	7	-	88–90
Goldman 2011	manual	AE	S	-	-	62	79
Goldman 2011	manual	French	S	-	-	57	81
Wilbanks 2015	manual	Spanish	S	-	15	68	79
Hosom 2009	manual	AE	R	-	-	82	94
	Lower bound in literature			n/a	15	57	79
	Upper bound in literature			n/a	7	82	94

Table 1. Schedule of the benchmarks set in the literature according to the four most popular measurements (Abbreviations: AE American English; BE British English; S spontaneous speech; R read-aloud speech; ms milliseconds; pct percentage).

MacKenzie and Turton (2020) later tested FAVE-Align on read-aloud and spontaneous British English and found 83 percent of read-aloud phones to fall within 20 milliseconds of the manual benchmark with a mean displacement of 8 milliseconds. They found between 76 percent and 90 percent of spontaneous boundaries to fall within 20 milliseconds of the manual benchmark with a mean displacement ranging between 8 and 20 milliseconds (2020, 9). Neither median errors nor 10-millisecond performance metrics were calculated.

Goldman (2011), in his development of EasyAlign for Praat, tested its accuracy on two fifteen-minute excerpts of spontaneous English and French speech. He compared performance against the alignments of two manual transcribers. For English, 50 percent and 51 percent of automatic alignments fell within 10 milliseconds of the standards set by human aligners 1 and 2, respectively; 77 percent and 75 percent fell within 20 milliseconds. For French, 49 percent and 52 percent of automatic alignments fell within 10 milliseconds of the standards set by human aligners 1 and 2, respectively; 79 percent and 82 percent fell within 20 milliseconds.

Wilbanks (2015) built a forced aligner for Spanish (*FASE*) that attained a 45 percent agreement rate for the 10 millisecond range and 70 percent for the 20 millisecond range. Mean differences between FASE and human alignment was 21 milliseconds.

Lastly, Hosom (2009) developed his own aligner for read-aloud English that is the sole aligner to come close to the standards set by FAVE-Align; namely, 80 percent within 10 milliseconds and 93 percent within 20 milliseconds of his manual alignments (2009, 364). Important, however, is that the standards set by FAVE-Align are based on spontaneous speech whereas Hosom's (2009) metrics are from read-aloud speech within the TIMIT corpus.

The trend between read-aloud speech and spontaneous speech is not at all as consistent as one would have thought; in other words, the popular aligners have not always fared better on read-aloud speech. Therefore, we have decided to consolidate both speech registers for the ensuing synopsis on benchmarks: The lower bounds in the literature on automatic alignment for (1) median onset difference from manual alignments, (2) mean onset difference from manual alignments, (3) the percentage of boundaries that fall within 10 milliseconds of the manual alignment, and (4) the percentage of boundaries that fall within 20 milliseconds of the manual alignment are 13, 31, 36 percent, and 57 percent, respectively. The upper bounds in the literature are 12, 8, 80 percent, and 94 percent, respectively.

2.4.2 Benchmarks for manual alignment

Focusing on the accuracy of automatic aligners can lead one to forget that human alignment can have its share of errors as well. Table 1 contains a schedule of the benchmarks set by the literature on manual alignment. Cosi et al. (1991) compared three manual alignments of spontaneous speech against a fourth 'gold-standard' reference. They found mean variation to be 7 milliseconds and that the poorest agreement rate was 88 percent and the highest agreement rate was 90 percent when the tolerance range was 20 milliseconds (1991, 694). Hosom (2009), who tested his

own alignments against the TIMIT corpus alignment, had an agreement rate of 82 percent for a tolerance of 10 milliseconds and 94 percent for a tolerance of 20 milliseconds. Goldman (2011) found human-to-human agreement for North American English to be 62 percent within the 10-millisecond range and 79 percent within the 20-millisecond range. For French, it was 57 percent and 81 percent, respectively. For Spanish, Wilbanks (2015) found human-to-human agreement to be 68 percent and 79 percent for the 10 and 20-millisecond thresholds, respectively. The mean difference in boundary placement between the two human aligners was 15 milliseconds.

In summary, the lower bounds in the literature on manual alignment for (1) median onset difference from manual alignments, (2) mean onset difference from manual alignments, (3) the percentage of boundaries that fall within 10 milliseconds of the manual alignment, and (4) the percentage of boundaries that fall within 20 milliseconds of the manual alignment are n/a, 15, 57 percent, and 79 percent, respectively. The upper bounds in the literature are n/a, 7, 82 percent, and 94 percent, respectively.

Evident here is that the lower bounds certainly exceed those offered by forced-alignment software but that the upper bounds are nearly identical. This is to say that the current technology is relatively mature, which implies that a lot can be gained by expanding it to lesser-studied languages. In the subsequent sections, the procedure for building the aligner will be discussed, and its performance will be assessed according to the minimal and maximal standards established the literature. The minimal standards will be taken from the lower bounds set by the literature on forced alignment (13, 31, 36 percent, 57 percent). The the maximal standards will be taken from the upper bounds set by the literature on both manual alignments and forced alignment, whichever of the metrics is superior (7, 8, 82 percent, 94 percent).

3. The current study

The present study adapted FAVE-Align to Swedish (henceforeth *SweFA*, the acronym for *Forced Alignment of Swedish*) and tested SweFA on the speech of nine adult male speakers of Stockholm Swedish. First, the acoustic models in FAVE-Align were relabeled according to their closest corresponding Swedish phoneme. Second, a Swedish pronunciation dictionary was procured and configured to the requirements set by the FAVE-Align and HTK architecture. Third, performance was tested on spontaneous and read-aloud speech from the aforementioned nine speakers. The following three sections outline the procedure in detail.

3.1 Adapting FAVE-Align to Swedish

FAVE-Align has transparent architecture and ample documentation, which makes it particularly handy for adaptation. Although is has been adapted before (DiCanio et al., 2013; Coto-Solano and Solórzano, 2017; Coto-Solano et al., 2018), detailed instructions for doing so have never been shared, which has resulted in an unfortunate stream of duplicated and uncoordinated efforts. The first author, who has ex-

Consonants						Vowels					
SweFAbet	IPA	grapheme	Swedish lexical example	closest English phoneme	ARPAbet	SweFAbet	IPA	grapheme	Swedish lexical example	closest English phoneme	ARPAbet
P	p	p	pil	p	P	II	iː	i	DIS	i	IY
B	b	b	bil	b	B	YY	yː	y	TYP	i	IY
T	t	t	tal	t	T	UU	ʉː	u	LUS	u	UW
D	d	d	dal	d	D	EE	eː	e	LETA	i	IY
K	k	k	kal	k	K	OE	øː	ö	SÖT	ʊ	UH
G	g	g	gas	g	G	OEE	œː	ö(+r)	DÖR	ʊ	UH
M	m	m	mil	m	M	AE	ɛː	ä	NÄT	ɛ	EH
N	n	n	nål	n	N	AEE	æː	ä(+r)	LÄR	æ	AE
NG	ŋ	ng, gn	ring	ŋ	NG	OO	uː	o	SOT	u	UW
R	r	r	ris	ɾ	D[7]	OA	oː	å	LÅS	oʊ	OW
F	f	f	fil	f	F	AA	ɑː	a	LAT	ɔ	AO
V	v	v	vår	v	V	IH	ɪ	i	DISK	ɪ	IH
TH	θ	th	thriller	θ	TH	YH	ʏ	y	FLYTTA	ɪ	IH
DH	ð	th	*that's it!*	ð	DH	EH	ɛ	e	LETT	ɛ	EH
S	s	s	sil	s	S	OEH	œ	ö	DÖRR	ʊ	UH
Z	z	z	guzz	z	Z	AEH	ɛ	ä	SÄRK	ɛ	EH
TJ	ɕ	tj	tjock	ʃ	SH	OH	ʊ	o	ROTT	ʊ	UH
SJ	ɧ	sj, sk, stj	sjuk	h	HH	UH	ɵ	u	LUDD	ʊ	UH
HH	h	h	hal	h	HH	OAH	ɔ	å	LOTT	ɔ	AO
J	j	j	jag	j	Y	AH	a	a	LASS	a	AA
L	l	l	lös	l	L	AJ	aj	aj	fajta	aɪ	AY
JH	dʒ	g, j	Jaffar	dʒ	JH	OJ	oj	oj	*oj!*	ɔɪ	OY
W	w	w	*walla!*	w	W	EJ	ɛj	ej	mejl	eɪ	EY
CH	tʃ	c, ch	cok	tʃ	CH	EU	ɛ͡ʉ	eu	euro	ɛ	EH
RT	ʈ	rt	fart	t	T	AU	a͡u	au	power	aʊ	AW
RD	ɖ	rd	bord	d	D						
RN	ɳ	rn	barn	n	N						
RS	ʂ	rs	fors	ʃ	SH						
RL	ɭ	rl	Karl	l	L						

Table 2. SweFAbet, corresponding IPA, grapheme, Swedish lexical example[8] , and closest English phoneme with ARPAbet.

pertise in Swedish phonetics, therefore scoured the code and identified change spots that would allow the use of the English HTK models for the closest corresponding Swedish phoneme. The second author, a seasoned programmer, proofed these change spots and made the hardcoded changes more pythonic. The original English monophones in FAVE-Align are done in *ARPAbet*, which is an ASCII-compatible system created by the Advanced Research Projects Agency's (ARPA) Speech Understanding Project. We created a similar system for Swedish that we refer to here as *SweFAbet*.

Table 2 provides a list of the Swedish phoneme inventory. The first column contains the SweFAbet monophone, the second column the corresponding IPA symbol, the third column the most common corresponding grapheme, the fourth column a Swedish lexical example (some are loanwords; e.g., *cok*), the fifth column the closest English phoneme, and the sixth column the ARPAbet monophone for that closest English phoneme.

The closest phoneme match was subjectively determined by the first author, and no testing was conducted to assess which phoneme would be more suitable. For example, Central Swedish NÄT falls between American English TRAP and DRESS, so we decided arbitrarily on DRESS (Arpabet EH). For Central Swedish /ɧ/, there are strong arguments for both selecting American /h/ and /ʃ/, so we decided arbitrarily on /h/. It is because of this process that we have referred to our adaptation as 'crude' and 'rapid'. Testing and optimizing phoneme matches would contradict the original aim of rapid prototype adaptation.

These adaptations are made in just six different locations within the FAVE-Align code. Since one aim of this paper is to be a resource for other researchers who wish to build a similar rapid prototype, detailed instructions on how we did this are provided in the Appendix.

After we programmed these substitutions in, we subsequently also built a prototype for Danish (not discussed in this paper), and the second author built a unicode-8 converter, an IPA converter, and a language-general shell to host the Danish, English, and Swedish aligners within one program (*LG-FAVE*, Young and McGarrah 2017). The program is free and accessible at https://github.com/mcgarrah/LG-FAVE.

3.2 Procuring and adapting a pronunciation dictionary

The larger project for which this Swedish adaptation was built required a comprehensive dictionary (Young, 2019), and two resources were particularly suitable for adaptation to the FAVE-Align format – *Folkets Lexikon* and *The NTS pronunciation dictionary*. These would not have been necessary for an experimental project that used, for example, a limited number of read-aloud sentences (e.g., Chodroff and Wilson 2017). Nonetheless, we have decided to dedicate some space here to the procurement of our dictionary because, as argued in Section 2.3, 75 percent of the time saved in automatic transcription comes from having a pronunciation dictionary that is both comprehensive and accurate. Furthermore, *The NTS pronunciation dictionary* is also publicly available for Danish and Eastern Norwegian ('bokmål'), so our adaptation can serve as a guideline for those who wish to duplicate SweFA for

these varieties.

3.2.1 Folkets Lexikon

Many proprietary dictionaries of Swedish are actually interface improvements to the Folkets Lexikon (*People's Lexicon*, Kann 2010; Kann and Hollman 2011), a state-funded project that sought to offer the first web-accessible Swedish dictionary. It was first published in 2009 and has undergone successive improvements through 2014.

The first author of this paper wrote a series of regular expressions to transform its XDXF format into the FAVE-compatible format. Although Folkets Lexikon has 200 000 total entries, it only has 18 928 pronunciation entries, which made it insufficient for spontaneous speech recordings.

3.2.2 NTS pronunciation dictionary for Swedish

In 2003, Nordic Language Technology Holdings, Inc (NTS) filed for bankruptcy, and the Norwegian National Library procured its intellectual property for public dissemination. At the time, NTS was working on Automatic Speech Recognition (ASR) for Danish, Eastern Norwegian ('bokmål'), and Central Swedish. All three of the NTS pronunciation dictionaries have been released to the public, but their accuracy could not be guaranteed. Whereas Folkets Lexikon is widely accepted as a credible source and has a chain of provenance in terms of its development, NTS is simply a file the first author "stumbled across". We therefore reconciled the NTS entries with Folkets Lexikon. The NTS entries matched all but 609 of the 18 928 Folkets Lexikon entries.

We therefore decided to use the NTS dictionary, converting its original IBM format into the FAVE-compatible format. The resulting product was a pronunciation dictionary with 927 167 entries. We also added approximately 3 000 slang words and programmed in a series of alternate elided pronunciations. Such alternate pronunciations are showcased in Figure 2 with the entry for '*cirkusartist*'. The final version of the dictionary has just over 16 million pronunciation entries.

3.3 Testing SweFA's performance

SweFA was tested on the speech of nine adult male speakers of Stockholm Swedish. Three of them speak the received Stockholmian variety, which is closest to what Riad (2014) refers to as *Central Swedish* (*centralsvenska*); three speak the traditional working-class variety, sometimes referred to as *Södersnack* or *Ekensnack* (Kotsinas, 1988b); three speak Stockholm's multiethnolect, sometimes referred to as *Rinkeby Swedish* of *Suburban Swedish* (*förortssvenska*) (Kotsinas, 1988a; Young, 2018).

The geographic origin of the speakers is plotted in the map in Figure 4. The map includes Stockholm's metro system because this is the spatial framework to which the city's residents typically associate its social dialects (Bijvoet and Fraurud, 2012). One might hear the comment 'he sounds very blue line' as a reference to Stockholm's multiethnolect. Likewise, one might hear 'that's so green line' in reference to the habitus of the white working class. In this study, speakers of the received variety hail

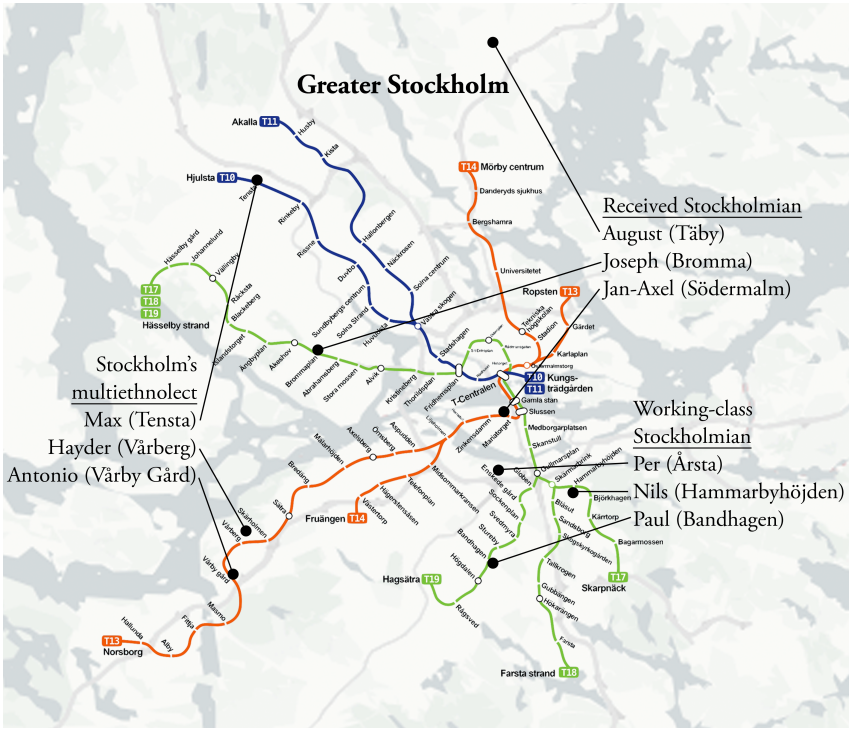

Figure 4. Map of greater Stockholm and its metro. Home neighborhoods of the nine speakers are plotted, and speakers are itemized according to their respective social dialects.

from the Center City and affluent suburbs. Speakers of working-class Stockholmian hail from the traditional white working-class strongholds in the Southeast. Speakers of Stockholm's multiethnolect hail from the multiethnic suburbs in the Northwest and Southwest.

Two speech styles were recorded for each of the nine speakers: spontaneous and reading. Both styles were taken from sociolinguistic interviews conducted by the first author. Criteria for treatment as 'spontaneous' were the presence of swearing, *channel cues* (Labov, 1972, 113) and/or a topic that was engaging for the speaker such as danger of death or supernatural occurrences (Labov, 1972). The reading task occurred at the end of the interview whereby the participant was asked to read an adaptation of *Cirkusen*, a speech-pathology diagnostic passage that contains multiple exemplars of every Swedish phoneme and tone accent (Morris and Zetterman, 2011).

Recordings were made on individual Zoom H1 recorders with self-powered Audio-Technica lavalier microphones in a quiet setting with minimal background noise. They were recorded in WAV format, mono, with a sample rate of 16 000 Hertz. The speech material was orthographically transcribed by native-language transcribers, financed by a grant from the Sven och Dagmar Salén Foundation. The transcriptions were then checked by the first author and subsequently phonetically time-aligned using SweFA. The first author then manually corrected the segmentations in accordance with standard segmentation protocol and the guidelines provided

in Engstrand et al. (2000). Manual correction of the alignments took an average of 68 seconds per recorded second (something that we discuss in Section 4). Segmental metrics were extracted using a customized adaptation of Brato's (2015) script for Praat (Boersma and Weenink, 2017).

For the spontaneous samples, pauses and hesitation markers were removed, and the first 1000 boundaries were measured. For the reading samples, the entire recording was measured after pauses and hesitation markers were removed, resulting in a range between 954 and 1040 boundaries.

4. Results

Table 3 organizes the nine speakers and two speech styles according to the four selected metrics. It also offers the minimal standards taken from the literature on forced alignment and the maximal standards derived from the literature on both forced alignment and manual alignment (reviewed in Section 2.4). The results that exceed the minimal standards are highlighted in light gray. The results that exceed the maximum standards are highlighted in dark gray.

For read-aloud speech, SweFA exceeds the minimal standard across all four metrics for every speaker and variety. It also exceeds the maximal standards for two speakers on the parameters of median onset difference: 5 for Jan-Axel and 6 for Hayder. For all of the spontaneous speech excerpts, SweFA satisfies at least one minimal benchmark. For five of the eight spontaneous speech excerpts, all minimal benchmarks are satisfied (August, Joseph, Nils, Hayder, Max).

When all speakers were consolidated and assessed as a whole – shown in the bottom row of Table 3, spontaneous speech exceeded three of the four minimal benchmarks, and read-aloud speech exceeded all four of the minimal benchmarks. For read-aloud speech, median and mean variation from the manual standard was 10 and 18 milliseconds, respectively. For the 10 and 20-millisecond tolerance range, 50 percent and 73 percent of alignments fell within range, respectively. While spontaneous speech performed less well, it still showed an accuracy level that is competitive with other aligners reported in the literature. Median and mean variation from the manual standard were 13 and 32 milliseconds, respectively. For the 10 and 20-millisecond tolerance range, 37 percent and 65 percent of alignments fell within range, respectively.

As disclosed in Section 3.3, manual correction of the alignments took us an average of 68 seconds per recorded second. The original orthographic transcriptions had an approximate ratio of 10:1, which meant that the final productivity ratio for human correction was 78:1.

As discussed in the closing of Section 2.2, the present study follows a series of other untrained prototypes for lesser-studied languages, including read-aloud Yoloxóchitl Mixtec (DiCanio et al., 2013), spontaneous Bribri (Coto-Solano and Solórzano, 2017), Cook Islands Maori (Coto-Solano et al., 2018), and read-aloud and spontaneous Baura, Bora, Even, and Sri Lankan Malay (Strunk et al., 2014). It is not possible to compare the accuracy of SweFA with the adaptations to Bribri

	Median onset difference	Mean onset difference	Pct 10 ms or less	Pct 20 ms or less
Benchmark lower bound	13	31	36%	57%
Benchmark upper bound	7	8	82%	94%

		Spontaneous					Read-aloud				
Stockholmian sociolect	Pseudonym	*n* boundaries	Median onset difference	Mean onset difference	Pct 10 ms or less	Pct 20 ms or less	*n* boundaries	Median onset difference	Mean onset difference	Pct 10 ms or less	Pct 20 ms or less
Received (*centralsvenska*)	August	1000	13	21	39%	70%	1001	13	22	37%	66%
	Joseph	1000	13	28	40%	69%	1026	10	27	52%	71%
	Jan-Axel	1000	16	78	35%	60%	1040	5	17	54%	74%
Working-class (*ekensnack*)	Per	1000	15	36	37%	62%	991	9	14	52%	75%
	Nils	1000	13	24	39%	66%	1012	12	18	42%	69%
	Paul	1000	14	33	33%	61%	954	10	16	51%	75%
Multiethnolect (*förortssvenska*)	Max	1000	12	19	42%	71%	1041	8	14	55%	76%
	Hayder	1000	13	22	36%	65%	1033	6	14	57%	77%
	Antonio	1000	14	27	35%	63%	1025	10	17	50%	72%
	all	9000	13	32	37%	65%	9123	10	18	50%	73%

Table 3. (top) Upper and lower performance standards from the literature. (bottom) Performance of SweFA for three male speakers of Stockholm's three main sociolects each in two speech styles according to four metrics. Results highlighted in light gray exceed the lowest standards in the literature; results highlighted in dark gray exceed the highest standards in the literature.

and Maori because they used different metrics. DiCanio et al. (2013, 2239), however, reported 32.3 percent and 52.3 percent of their (read-aloud) alignments falling within 10 and 20 milliseconds of the manual benchmarks, respectively, in contrast to the 50 percent and 73 percent reported here. Strunk et al. (2014, 3944) reported a median variation from the manual standard that ranged between 30 (read-aloud Bora) and 160 (spontaneous Bora) milliseconds in contrast to 10–13 reported here. Mean variation from the manual standard fell between 148 (read-aloud Bora) and 290 (spontaneous Bora) milliseconds in contrast to 18–32 reported here.

During the peer-review process it was pointed out that it is difficult to decide whether to attribute the success of the aligner to an excellent dictionary or to the typological similarity between Swedish and English. Recall that we reported in Section 3.2.2 that we added a high number of *elided* pronunciation options, bringing the entry number up from 1 million to 16 million entries (exemplified in Figure 2). To separate these two factors, we conducted a pilot analysis in which we ran the aligner using the "unexpanded" dictionary on the spontaneous speech of August, Paul and Max. We found only marginal differences. Referring back to Table 3, August showed 13, 21, 39% and 70% for median onset difference, mean onset difference, percentage of onsets that fell within 10 milliseconds of the manual benchmark, and percentage of onsets that fell within 20 milliseconds of the manual benchmark, respectively, with the expanded dictionary. With the "unexpanded" dictionary, these figures were 13, 22, 37% ,and 68%, respectively. For Paul, the "expanded" metrics in Table 3 were 14, 33, 33%, and 61%, and the "unexpanded" metrics were 16, 35, 32% and 59%, respectively. For Max, the "expanded" metrics in Table 3 were 12, 19, 42%, 71%, and the "unexpanded" metrics were 12, 20, 43%, and 71%. We did not conduct this comparison for all 18 speech samples, but we believe this post hoc analysis buttresses the conclusion that the aligner's success is mostly due to the typological similarity between English and Swedish.

5. Conclusion

SweFA, our Swedish adaptation of FAVE-Align, aligns the three main varieties of Stockholm Swedish at a competitive level of accuracy according to the minimal benchmarks set by the literature. This is of course important for Swedish phonetics research, but the broader implication is that researchers of other Nordic languages can rapidly adapt a prototype from FAVE-Align and expect a rewarding return on the endeavor. This is especially the case for read-aloud speech, where all nine test samples met all four benchmarks separately and as a whole.

As it pertains to the aligner working better on a particular Stockholmian variety, no significant trend emerged; rather, the variation appears to be idiolectal. For example, the spontaneous speech of Jan-Axel, Paul and Antonio performed similarly according to the 10ms and 20ms metrics. While their respective varieties are quite different, all three speakers mumble and have substantial vocal fry in their speech, which may be the reason behind SweFA's hindered effectiveness.

For researchers who have little interest in chancing their analyses on *pure* au-

tomatic alignment, manually correcting the alignments from an automatic prototype like SweFA can cut time spent by a half. This is to say that even if a study required manual alignments, using our proposed prototype as a starting point would still result in considerable time savings. As we disclosed in Section 2.3, our own manual capacity was 232:1, which meant that manually aligning the current 1899-second dataset would have taken approximately 122 hours ($\frac{232 \cdot 1899}{3600}$). With SweFA, our manual corrections took 41 hours ($\frac{78 \cdot 1899}{3600}$). The actual adaptation of FAVE-Align took about 4 hours, and the base adaptation of the NTS dictionary took another 8 hours. This translates into a time savings of 69 hours for this project.

A serious hurdle for aligning a lesser-studied language is procuring a sufficiently comprehensive pronunciation dictionary. While such dictionaries also exist for Danish and Eastern Norwegian by means of the NTS archives, they are lacking for other Nordic languages. For those remaining languages, the adaptation proposed here is particularly valuable for laboratory investigations that require a finite number of read-aloud sentences to be aligned (as opposed to open-ended spontaneous speech).

We conclude by proposing that phonetic investigations of the Nordic languages could benefit from 'untrained' aligners such as SweFA. Whereas prior untrained models have usually rendered poor results, SweFA's alignment of Swedish is as accurate as many custom-trained aligners of English. The implication here is of course that FAVE-Align is more easily adaptable to a language typologically closer to English than, for example, Finnish or Sami[9] . As we indicated earlier in the paper, Swedish has become somewhat underrepresented in the contemporary Phonetics literature. This is similarly the case for Danish and Eastern Norwegian, and of course the many other understudied Nordic varieties like Estonian Swedish, Faroese, Fenno-Swedish, Icelandic, the Northern and Western Norwegian dialects, and Övdalian. Our hope is that phoneticians can use our template to reduce the resource intensity of their future research endeavors.

APPENDIX

Detailed instructions for Adapting FAVE-Align to Swedish

The most recent version of FAVE-Align is downloadable from https://github.com/JoFrhwld/FAVE. Similarly, instructions on how to use it and how/where to download HTK and SoX are at https://github.com/JoFrhwld/FAVE/wiki/FAVE-align.

FAVE-Align was built using Python. Before any of the below steps are initiated, *be sure that you have installed FAVE-Align properly and that you have executed it successfully for English*. That way, if you encounter any problems in the below steps, you know it is because of your changes and not because of some other pre-existing bug.

Before adapting the software, the monophones that your language will use need to be defined and coded with ASCII characters. The ASCII requirement cannot be changed in FAVE-Align because the limitations are set by HTK, which is proprietary and encrypted. The closest corresponding American-English sound should be mapped to it as is shown for SweFAbet in Table 2. This mapping should be done

subjectively and to the best of what you know about the language of study (fortunately, FAVE is quite forgiving). Your pronunciation dictionary must use these same monophones.

In order to repurpose the English acoustic models over to the SweFAbet inventory, six files within the Folder entitled `FAVE-align` must be altered:

1. `/FAVE-align/FAAValign.py`
2. `/FAVE-align/model/monophones`
3. `/FAVE-align/model/16000/hmmdefs`
4. `/FAVE-align/model/8000/hmmdefs`
5. `/FAVE-align/model/11025/hmmdefs`
6. `/FAVE-align/model/dict`

Step 1: Adapt `/FAVE-align/FAAValign.py`

Figure A1 shows the English monophones on lines 97 and 98 of the original FAAValign.py script.

```
97  CONSONANTS = ['B', 'CH', 'D', 'DH','F', 'G', 'HH', 'JH', 'K', 'L', 'M', 'N', 'NG', 'P', 'R', 'S',
        ↪ 'SH', 'T', 'TH', 'V', 'W', 'Y', 'Z', 'ZH']
98  VOWELS = ['AA', 'AE', 'AH', 'AO', 'AW', 'AY', 'EH', 'ER', 'EY', 'IH', 'IY', 'OW', 'OY', 'UH', 'UW'
        ↪ ]
99  STYLE = ["style", "Style", "STYLE"]
100 STYLE_ENTRIES = ["R", "N", "L", "G", "S", "K", "T", "C", "WL", "MP", "SD", "RP"]
```

Figure A1. Section of FAVE's Python code that defines monophones

The English monophones on lines 97 and 98 need to be replaced with the monophones of the new language. The new monophones must be ASCII-compatible. Most of the monophones are phonemes, but some are allophones and diphthongs (like AJ or OJ below). The new SweFAbet monophones are entered into lines 97 and 98 in Figure A2.

```
97  CONSONANTS = ['B', 'CH', 'D', 'DH','F', 'G', 'HH', 'J', 'JH', 'K', 'L', 'M', 'N', 'NG', 'P', 'R',
        ↪ 'RD', 'RL', 'RN', 'RS', 'RT', 'S', 'SJ', 'T', 'TH', 'TJ', 'V', 'W', 'Z', 'ZH']
98  VOWELS = ['AA', 'AE', 'AEE', 'AEH', 'AH', 'AJ', 'AU', 'EE', 'EH', 'EJ', 'ER', 'EU', 'IH', 'II', '
        ↪ OA', 'OAH', 'OE', 'OEE', 'OEH', 'OH', 'OJ', 'OO', 'UH', 'UU', 'YH', 'YY']
99  STYLE = ["style", "Style", "STYLE"]
100 STYLE_ENTRIES = ["R", "N", "L", "G", "S", "K", "T", "C", "WL", "MP", "SD", "RP"]
```

Figure A2. Section of SweFA's Python code that defines monophones

FAVE-Align measures stress on the vowel of each syllable, and this is coded with a 1 for primary stress, a 0 for no stress, and a 2 for secondary stress. Swedish, however, is a pitch-accent language (see Riad 2014), so Accent 1 is denoted with a 1 on the vowel, Accent 2 is denoted with a 3 on the vowel, secondary stress is denoted with a 2 on the vowel, and lack of stress is denoted with a 0 on the vowel.

Line 468 should be changed such that it can accommodate monophones longer than 3 ASCII characters as well as the additional stress codings 0, 1, 2, and 3. This is shown in Figure A3. Line 500 needs a small change as well, shown in Figure A4.

Even though FAVE-Align is restricted to ASCII, its script has a number of sophisticated protections built to keep things running even if there are non-ASCII characters in the transcription that would otherwise upset the program. These 'fixes', so

Before conversion

```
468    if not ((len(p) == 3 and p[-1] in ['0', '1', '2'] and p[:-1] in VOWELS) or (len(p) <= 2 and p
           ↪ in CONSONANTS)):
```

After conversion

```
468    if not ((len(p) <= 5 and p[-1] in ['0', '1', '2', '3'] and p[:-1] in VOWELS) or (len(p) <= 3
           ↪ and p in CONSONANTS)):
```

Figure A3. Section of SweFA's Python code that defines monophone string length and stress numbering

Before conversion

```
500    if len(w) > 3 and len(phones) < 2:
```

After conversion

```
500    if len(w) > 5 and len(phones) < 2:
```

Figure A4. Additional section of SweFA's Python code that defines monophone string length

to speak, begin on line 510, shown in Figure A5(a). As line 513 indicates, this only works for transcriptions in Unicode 8 (UTF-8). The regexes from lines 517 to 520 turn the four most-common rich-text single quotes into an ASCII single quote. The regexes from lines 521 to 524 turn the four most-common rich-text double quotes into an ASCII double quote.

Since the Swedish keyboard has two other types of double quotes, these were added to lines 525 and 526, shown in Figure A5(b). Since the Swedish characters Ä, ä, Ö, ö, Å, å are not ASCII-compatible, we selected $, $, #, #, @, and @, respectively, shown on lines 527 to 532 in Figure A5(b). Crucially, these were then also substituted into the pronunciation dictionary.

Step 2: Adapt `/FAVE-align/model/monophones`

This file contains a simple list of all of the monophones. Note, however, that the vowels must be listed with all possible numerical stress markings. In the case of the SweFA adaptation, this means 0 through 3.

Step 3: Adapt `/FAVE-align/model/16000/hmmdefs`

The hidden Markov vectors for each monophone-including-stress is defined in the `hmmdefs` files. Figure A6 shows a snapshot of the vectors for the monophone UH0 for a 16000-Hertz sound file (/FAVE-align/model/16000/hmmdefs). The monophone is defined in the quotes that follow `~h`. The Swedish monophones were substituted in for the closest-sounding English monophone, shown in Figure A7. Since there are more Swedish monophones than English, many of the vectors were duplicated. For example, since only three vectors exist for UH (UH0, UH1, and UH2), UH1 was duplicated and then one duplicate was changed to OEH1 and the other to OEH3. UH0 became OEH0, and UH2 became OEH2.

Steps 4/5: Adapt `/FAVE-align/model/8000/hmmdefs` and

Before conversion

```
# substitute any 'smart' quotes in the input file with the corresponding
# ASCII equivalents (otherwise they will be excluded as out-of-
# vocabulary with respect to the CMU pronouncing dictionary)
# WARNING: this function currently only works for UTF-8 input
def replace_smart_quotes(all_input):
  cleaned_lines = []
  for line in all_input:
    line = line.replace(u'\u2018', "'")
    line = line.replace(u'\u2019', "'")
    line = line.replace(u'\u201a', "'")
    line = line.replace(u'\u201b', "'")
    line = line.replace(u'\u201c', '"')
    line = line.replace(u'\u201d', '"')
    line = line.replace(u'\u201e', '"')
    line = line.replace(u'\u201f', '"')
    cleaned_lines.append(line)
  return cleaned_lines
```

After conversion

```
# substitute any 'smart' quotes in the input file with the corresponding
# ASCII equivalents (otherwise they will be excluded as out-of-
# vocabulary with respect to the CMU pronouncing dictionary)
# WARNING: this function currently only works for UTF-8 input
def replace_smart_quotes(all_input):
  cleaned_lines = []
  for line in all_input:
    line = line.replace(u'\u2018', "'")
    line = line.replace(u'\u2019', "'")
    line = line.replace(u'\u201a', "'")
    line = line.replace(u'\u201b', "'")
    line = line.replace(u'\u00B4', "'")
    line = line.replace(u'\u0060', "'")
    line = line.replace(u'\u201c', '"')
    line = line.replace(u'\u201d', '"')
    line = line.replace(u'\u201e', '"')
    line = line.replace(u'\u201f', '"')
    line = line.replace(u'\u00E4', '$')
    line = line.replace(u'\u00F6', '#')
    line = line.replace(u'\u00E5', '@')
    line = line.replace(u'\u00C4', '$')
    line = line.replace(u'\u00D6', '#')
    line = line.replace(u'\u00C5', '@')
    cleaned_lines.append(line)
  return cleaned_lines
```

Figure A5. Section of FAVE's Python code that converts potential UTF-8 characters in the transcription into ASCII

/FAVE-align/model/11025/hmmdefs

These are done exactly the same way as Step 3.

Step 6: Adapt /FAVE-align/model/dict

This file is the pronunciation dictionary. All entries must be in ASCII and sit on a separate line. In the case of Swedish, this meant substituting Ä, ä, Ö, ö, Å, å for $, $, #, #, @, and @, respectively. A space should separate the entry from its pronunciation, and a space must lie between every monophone in the pronunciation entry. Figure A8 shows an example.

ENDNOTES

1. For example, we do not rule out the possibility that various singular researchers may actually be in possession of a forced aligner for every Nordic language. However, the 'high investment/ low reward' of disseminating methodological improvements in our field may be disincentivizing the sharing of such innovations. It is, for example, still rare that one encounters methodological papers in peer-reviewed journals.
2. This claim has been confirmed in personal conversations with Nicolai Pharao and Gert Foget Hansen at Copenhagen University, Sverre Stausland Johnsen at Oslo University, Per Erik Solberg at the Norwegian National Library, and Johan Gross at Gothenburg Univer-

```
40470 ~h "UHO"
40471 <BEGINHMM>
40472 <NUMSTATES> 5
40473 <STATE> 2
40474 <NUMMIXES> 32
40475 <MIXTURE> 1 3.032761e-02
40476 <MEAN> 39
40477  -9.564137e-01 -7.124956e-02 5.896883e-01 -4.774669e-01 -1.047210e-01 1.102472e+00 3.488875e-01
       ↪ -6.569489e-01 -7.019844e-01 -1.973782e-01 8.310850e-01 8.163610e-03 8.529254e-01 2.679475
       ↪ e-01 8.599031e-02 1.631898e-02 -4.532797e-02 1.401701e-02 -3.564458e-01 -2.504274e-01
       ↪ 1.878152e-01 2.014271e-01 4.116456e-03 1.554149e-02 5.204271e-02 8.566935e-02 2.238926e
       ↪ -01 -6.333413e-02 -1.319931e-01 8.102211e-02 1.219049e-01 -1.757646e-01 -1.315476e-01
       ↪ 1.229765e-01 1.139365e-01 6.854153e-02 -1.452735e-01 3.484021e-02 -8.926608e-02
40478 <VARIANCE> 39
40479  7.919521e-02 1.469982e-03 8.185884e-03 1.287063e-02 1.599258e-02 5.763948e-02 8.481910e-02
       ↪ 1.419987e-03 2.595889e-02 2.760400e-02 8.337084e-03 8.251499e-02 1.550067e-03 3.672207e
       ↪ -02 1.828414e-03 8.989252e-03 4.794365e-03 7.035460e-03 5.657048e-02 1.096972e-02
       ↪ 1.880060e-02 6.692914e-03 9.765797e-04 1.582222e-02 3.686350e-03 7.117046e-03 2.351006e
       ↪ -03 1.668679e-04 4.574205e-05 1.275223e-07 9.643809e-05 1.635860e-03 2.870251e-04
       ↪ 2.580893e-04 7.450095e-04 2.130718e-04 1.514308e-04 7.179412e-04 5.170816e-05
40480 <GCONTS> -1.604137e+02
40481 <MIXTURE> 2 3.032777e-02

      ⋮

      ⋮

41052 <VARIANCE> 39
41053  2.158116e-02 1.137838e-01 7.635124e-02 6.735466e-03 1.194743e-01 4.087792e-01 2.257451e-02
       ↪ 6.661187e-03 2.487542e-02 9.915646e-04 5.491704e-03 1.015392e-02 7.131799e-02 1.064542e
       ↪ -02 9.411335e-05 1.085411e-03 8.868680e-03 9.618518e-03 5.244023e-04 4.481581e-03
       ↪ 6.476781e-03 1.827894e-03 3.509800e-04 1.184476e-03 1.716148e-03 1.448500e-02 1.441586e
       ↪ -03 5.106049e-03 1.013570e-03 2.609197e-03 1.523106e-03 1.076035e-03 3.263119e-04
       ↪ 2.520934e-05 2.206339e-03 9.860150e-04 3.037707e-03 1.619887e-04 1.385237e-05
41054 <GCONTS> -1.517275e+02
41055 <TRANSP> 5
41056  0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
41057  0.000000e+00 3.933974e-01 6.066025e-01 0.000000e+00 0.000000e+00
41058  0.000000e+00 0.000000e+00 6.226407e-01 3.773592e-01 0.000000e+00
41059  0.000000e+00 0.000000e+00 0.000000e+00 6.969731e-01 3.030269e-01
41060  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
41061 <ENDHMM>
41062 ~h "UH1"
41063 <BEGINHMM>
41064 <NUMSTATES> 5
41065 <STATE> 2
41066 <NUMMIXES> 32
41067 <MIXTURE> 1 2.262416e-02
41068 <MEAN> 39
```

Figure A6. Excerpt from lines 40470 to 41068 of the hidden Markov model vectors for the monophone UH in unstressed position (indicated by `~h "UH0"`)

Step 3A: Identify Original HMMdefs

```
40471 ~h "UH0"
40472 <BEGINHMM>
40473 <NUMSTATES> 5

      ⋮

41062 ~h "UH1"
41063 <BEGINHMM>
41064 <NUMSTATES> 5

      ⋮

41654 ~h "UH2"
41655 <BEGINHMM>
41656 <NUMSTATES> 5
```

Step 3B: Duplicate where necessary

```
40471 ~h "UH0"
40472 <BEGINHMM>
40473 <NUMSTATES> 5

      ⋮

41062 ~h "UH1"
41063 <BEGINHMM>
41064 <NUMSTATES> 5

      ⋮

42246 ~h "UH1"
42247 <BEGINHMM>
42248 <NUMSTATES> 5

      ⋮

42838 ~h "UH2"
42839 <BEGINHMM>
42840 <NUMSTATES> 5
```

Step 3C: Change monophones to new names

```
40471 ~h "OEH0"
40472 <BEGINHMM>
40473 <NUMSTATES> 5

      ⋮

41062 ~h "OEH1"
41063 <BEGINHMM>
41064 <NUMSTATES> 5

      ⋮

42246 ~h "OEH3"
42247 <BEGINHMM>
42248 <NUMSTATES> 5

      ⋮

42838 ~h "OEH2"
42839 <BEGINHMM>
42840 <NUMSTATES> 5
```

Figure A7. Converting the FAVE-Align vectors for UH to SweFA's OEH. First UH1 and UH2 are duplicated, then the names are changed.

```
PLANE  P L AA3 N EHO
PLANEKONOMI  P L AA3 N EHO K OHO N OHO M II2
PLANEKONOMIER  P L AA3 N EHO K OHO N  M II2 EHO R
PLANEKONOMIERNA  P L AA3 N EHO K OHO N OHO M II2 EHO  RN AHO
PLANEKONOMIERNAS  P L AA3 N EHO K OHO N OHO M II2  EHO  RN AHO S
PLANEKONOMIERS  P L AA3 N EHO K OHO N OHO  M II2  EHO RS
PLANEKONOMIN  P L AA3 N EHO K OHO N OHO M II2 N
PLANEKONOMINS  P L AA3 N EHO K OHO N OHO M II2 N S
PLANEKONOMIS  P L AA3 N EHO K OHO  N OHO M II2 S
PLANEKONOMISKA  P L AA3 N EHO K OHO N OA2 M IHO  S K AHO
PLANEN  P L AA1 N EHO N
```

Figure A8. Dictionary format for `/FAVE-align/model/dict`. Every entry requires its own line, the entry must be in ASCII, and the entry is separated from its pronunciation by a single space. Subsequent spaces separate monophones.

sity/University West.

3. We acknowledge, of course, that Swedish has *historically* played a seminal role in phonetics research (Jakobson et al., 1951; Fant, 1952).
4. Pitch accent 2 is dyadic, containing an initial fall or smaller peak on the stressed syllable followed by a delayed large peak on the post-tonic syllable. This is why the first "2" is written as a superscript and the second "2" as a subscript. I refer the reader to Riad (2014, 181) for similar conventions for denotating accents 1 and 2 in Swedish.
5. Translation: 'The circus was on its way. Their poster'
6. 400:1 is reported to be the typical upper limit (Yuan et al., 2013, 2306).
7. This is of course variable in Stockholm Swedish, and the decision for which ARPABET model to use is debatable. In the present corpus, we found that syllable-final /r/ was either completely elided or manifested as [ɻ]. The latter, however, was a rare occurrence because an overwhelming majority of syllable-final /r/ values become syllable onsets due to the sandhi effect in fluent speech (*han är ung*). As such, onset /r/ was quite frequent and manifested itself most often as a flap or retroflex flap (and very occasionally as a trill among multiethnolect speakers). In light of all of this, we ruled out using FAVE's R monophone for Swedish /r/. We were therefore left with the models for T or D. We felt, however, that the portion of aspirated t-allophones that contributed to the model for T ([tʰ] in *talk* vs. [ɾ] in *ate it*) would have made it less optimal than D ([d] in *dog* vs. [ɾ] in *made it*). Therefore, we settled on D.
8. For vowels, lexical sets are provided as established by the SweDia corpus (Engstrand et al., 2000) in small caps as established by Leinonen (2010).
9. For non-Germanic languages like Finnish or Sami, the papers on adapting FAVE to Bora or Yoloxóchitl Mixtec might offer a better guideline as to accuracy limits (Strunk et al., 2014; DiCanio et al., 2013).

REFERENCES

Bijvoet, E. and Fraurud, K. (2012). Studying high-level (L1–L2) development and use among young people in multilingual Stockholm. *Studies in Second Language Acquisition*, 34(2):291–319.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [Computer software], Version 6.0.36.

Brato, T. (2015). TB-Basic Vowel Analysis, Version 2.2 [Computer software].

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47.

Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., and de Souza, R. N. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, 75:113–147.

Cosi, P., Falavigna, D., and Omologo, M. (1991). A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In *Second European Conference on Speech Communication and Technology*, pages 693–696.

Coto-Solano, R., Nicholas, S. A., and Wray, S. (2018). Development of Natural Language Processing Tools for Cook Islands māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33.

Coto-Solano, R. and Solórzano, S. F. (2017). Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. *CLEI Electronic Journal*, 20(1):2–1.

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., and García, R. C. (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.

Dodsworth, R. and Benton, R. A. (2017). Social network cohesion and the retreat from Southern vowels in Raleigh. *Language in Society*, 46(3):371–405.

Engstrand, O., Bruce, G., Elert, C.-C., Eriksson, A., and Strangert, E. (2000). *Databearbetning i SweDia 2000: segmentering, transkription och taggning. Version 2.2* [*Data work in SweDia 2000: transcription, segmentation, and tagging*]. University of Gothenburg, Gothenburg.

Evanini, K. (2009). *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*. PhD thesis, Department of Linguistics, University of Pennsylvania.

Fant, G. (1952). Acoustic Analysis of Speech – A Study for the Swedish Language. *Ericsson Technics*.

Fromont, R. and Hay, J. (2012). LaBB-CAT: An annotation store. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 113–117.

Goldman, J.-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association*, pages 3233–3236.

Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Gross, J., Boyd, S., Leinonen, T., and Walker, J. A. (2016). A tale of two cities (and one vowel): Sociolinguistic variation in Swedish. *Language Variation and Change*, 28(2):225–247.

Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4):352–368.

Jakobson, R., Fant, C. G. M., and Halle, M. (1951). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. The MIT Press, Cambridge, MA.

Kann, V. (2010). KTHs morfologiska och lexikografiska verktyg och resurser. *LexicoNordica*, (17).

Kann, V. and Hollman, J. (2011). *Slutrapport för projektet Vidareutveckling av Folkets lexikon*.

Kisler, T., Reichel, U., Schiel, F., Draxler, C., Jackl, B., and Pörner, N. (2016). BAS Speech Science Web Services - an Update of Current Developments. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Kotsinas, U.-B. (1988a). Rinkebysvenska - en dialekt? [Rinkeby Swedish - a dialect?]. In Linell, P., Adelswärd, V., Nilsson, T., and Pettersson, P. A., editors, *Svenskans beskrivning 16* [*The description of Swedish 16*], volume 1, pages 264–278. Tema Kommunikation, Linköping.

Kotsinas, U.-B. (1988b). *Stockholmspråk i förändring* [*Stockholm's language in change*], volume 1, pages 133–147. Lund University Press, Lund.

Kotsinas, U.-B. (1994). Snobbar och pyjamastyper: Ungdomskultur, ungdomsspråk och gruppidentiteter i Stockholm [Snobs and Pyjama types: Youth culture, youth language and group identities in Stockholm]. In Fornäs, J., Boethius, U., Forsman, M., Ganetz, H., and Reimer, B., editors, *Ungdomskultur i Sverige* [*Youth Culture in Sweden*], pages 311–336. Brutus Östlings Bokförlag, Stockholm.

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

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

Leinonen, T. (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. Groningen Dissertations, Groningen.

Lindh, J. (2007). Semi-Automatic Aligning of Swedish Forensic Phonetic Phone Speech in Praat using Viterbi Recognition and HMM [Unpublished manuscript].

MacKenzie, L. and Turton, D. (2020). Assessing the accuracy of existing forced alignment

software on varieties of British English. *Linguistics Vanguard*, 6(1):1–14.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech*, pages 498–502.

Morris, U. and Zetterman, H. (2011). Från bondgård till cirkus. Konstruktion av en högläsningstext för bedömning av röst-och talfunktion och talandning [From farm to circus. Constructing a reading passage for assessment of voice and speech faculties]. Master's thesis, Department of Speech Therapy, Karolinska Institutet, Stockholm, Sweden.

Nordberg, B. (1975). Contemporary social variation as a stage in a long-term phonological change. In Dahlstedt, K.-H., editor, *The Nordic languages and modern linguistics 2*. Almqvist & Wiksell, Stockholm.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society, Piscataway.

Riad, T. (2014). *The Phonology of Swedish*. Oxford University Press, Oxford.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) Program Suite* [*Computer software*].

Schultz, T. and Schlippe, T. (2014). GlobalPhone: Pronunciation Dictionaries in 20 Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 337–341, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Strunk, J., Schiel, F., Seifart, F., et al. (2014). Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMaUs. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation: May 26-31, 2014*, pages 3940–3947, Paris. European Language Resources Association (ELRA).

Thomas, E. and Carter, P. (2006). Prosodic rhythm and African American English. *English World-Wide*, 27:331–355.

Torgersen, E. and Szakay, A. (2012). An investigation of speech rhythm in London English. *Lingua*, 122(7):822–840.

Wilbanks, E. (2015). The development of FASE (Forced Alignment System for Español) and implications for sociolinguistic research. *paper presented at New Ways of Analyzing Variation 44, University of Toronto, 22–25 October*.

Young, N. (2018). Talrytmens sociala betydelse i det senmoderna Stockholm [The social meaning of speech rhythm in late-modern Stockholm]. *Nordand - Nordic Journal of Bilingualism Research*, 13(1):41–63.

Young, N. (2019). *Rhythm in late-modern Stockholm – Social stratification and stylistic variation in the speech of men*. PhD thesis, Department of Linguistics, Queen Mary, University of London.

Young, N. and McGarrah, M. (2017). Introducing NordfA - Forced Alignment of Nordic Languages. *Presentation at New Ways of Analyzing Variation (NWAV) 46, Madison, USA*.

Young, S. J., Woodland, P. C., and Byrne, W. J. (1993). *HTK: Hidden Markov Model Toolkit V1. 5*. Cambridge Univ. Eng. Dept. Speech Group and Entropic Research Lab. Inc., Cambridge.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics 2008*, pages 5687–5690.

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., and Wang, W. (2013). Automatic phonetic segmentation using boundary models. In Bimbot, F., Cerisara, C., Fougeron,

C., Gravier, G., Lamel, L., Pellegrino, F., and Perrier, P., editors, *INTERSPEECH 2013, Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29*, pages 2306–2310.