⚙Autophon user guide
Swedish
Model: SweFA version 2.0

1 Introducing Autophon and forced alignment

Autophon is a **free** online forced aligner. *Forced alignment* (FA) refers to the automatic process by which speech recordings are phonetically time-stamped with the help of Hidden Markov models or Deep Neural Networks. Autophon uses the latter by means of the *Montreal Forced Aligner*[1], which is built on the Kaldi toolkit[2]. The app outputs a time-stamped phonetic annotation, readable in Praat (Boersma and Weenink 2017), that is based on an optimization of two user inputs: (1) the speech recording and (2) a corresponding orthographic transcription.

Forced alignment is important because it automates something that is resource-intensive when done manually. A typical phonetic annotation can take between 250 and 400 minutes per recorded minute. In a place like Scandinavia – where labor costs are high – this cost has presented a barrier for linguists.

For a forced alignment tool to work, an acoustic model must be trained on the specific language, and an accompanying pronunciation lexicon must be built that covers every word in the language (See section 5).

Numerous forced aligners are in circulation and available to download and use. However, they often are command-line based and rely on operating systems (OS) that may be outdated and/or incompatible with your OS. Therefore, *Autophon aims to offer an **OS-agnostic** and **user-friendly** option for phoneticians around the world.*

2 Using the app

Logging in You can create an account by following the relatively intuitive guidelines on our website. We require an account because we wish to keep track of usage in order to make a case for funders. Furthermore, an open system makes us vulnerable to bot attacks. After registering, a verification email will be sent to you with a link that you must click on to verify your account. If you do not receive the email, first check your spambox and then wait at least 15 minutes before contacting tech support.

Cost Autophon is free of charge.

Adding files First go to the *Aligner* tab and click *Add files*. A box will appear with the heading *Transcription Mode: change transcription mode*. Click on the heading to select one of four *Transcription Modes*. Once your transcription mode has been selected, use the file browser underneath to select your files.

Transcription modes Four different *transcription modes* are available, each named according to the field in which the format is most common: *Experimental Linguistics A*, *Experimental Linguistics B*, *Computational Linguistics*, and *Variationist Linguistics*. Each can be selected by clicking on one of the boxes illustrated in Figure 1. The boxes illustrate a typical file structure for each mode and provide a link to a video that offers detailed formatting instructions.

Experimental linguistics A: In this mode, you upload a master transcription spreadsheet along with corresponding audio files — one by one or within a zip file. The master sheet should have two columns: column 1 holds the audio file names in your folder; column 2 holds the corresponding transcriptions. This format is similar to that used by, e.g., CommonVoice[3] and assumes that each audio file contains a short snippet of speech, which means that time stamps are *not* permitted. If you have a master transcription spreadsheet with time stamps, you are in the wrong transcription setting and need to select *Experimental Linguistics B*, described below. The master transcription sheet can be in a two-column Excel `xlsx` or tab-delimited file with either the extensions `txt` or `tsv`.

Experimental linguistics B: In this mode, you also upload a master transcription spreadsheet with corresponding audio files— one by one or within a zip file. Unlike in mode A, it should have four columns: column 1 holds the names of the audio files in your folder; column 2 — start times; column 3 — end times; column 4 — transcription. This mode is designed for longer audio files that warrant multiple

---

[1]McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger (2017)

[2]Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlicek, Qian, Schwarz, et al. (2011)

[3]https://commonvoice.mozilla.org

Experimental Ling A

(click to see video guide)

```
yourzip.zip
├ yourtrans.xlsx/tsv/txt
├ file0001.wav
├ file0002.wav
├ file0003.wav
├ …
└ file9999.wav
```

Transcriptions in a master file absent of time stamps - as separate rows with separate audio* files for each transcription.

Experimental Ling B

(click to see video guide)

```
yourzip.zip
├ yourtrans.xlsx/tsv/txt
├ file01.wav
├ file02.wav
├ file03.wav
├ …
└ file99.wav
```

Transcriptions in a master file with start and end time stamps with more than one row per audio* file.

Computational Ling

(click to see video guide)

```
yourzip.zip
├ file0001.lab
├ file0001.wav
├ file0002.lab
├ file0002.wav
├ file0003.lab
├ file0003.wav
├ …
├ file9999.lab
└ file9999.wav
```

Transcriptions as separate same-name lab and audio* files, absent of time stamps.

Variationist Ling

(click to see video guide)

```
yourzip.zip
├ file01.TextGrid
├ file01.wav
├ file02.eaf
├ file02.wav
├ file03.tsv
├ file03.wav
├ file04.xlsx
├ file04.wav
├ …
├ file99.txt
└ file99.wav
```

Longer transcription files in TextGrid, eaf, tsv, txt, or xlsx format with same-name audio* files.

Figure 1: The Transcription Mode selection menu for Autophon.

Figure 2: Autophon will output the finished TextGrids using an identical subfolder structure as the uploaded file.

lines of transcription. The master transcription sheet should either be a four-column Excel `xlsx` or a tab-delimited file with either the extensions `txt` or `tsv`. Time stamps must be in *seconds formatted as real numbers*. European comma decimals are accepted (e.g., **1,23**) as well as Anglo-American period decimals (e.g., **1.23**). What will not work, however, are time stamps with colons, minutes, or hours (e.g., **00:00:01.23**).

Computational linguistics: In this mode, you upload pairs of `lab` and audio files by the same name — one by one or within a zip file. These so-called `lab` files are simply text files that contain a single transcription phrase that matches the speech within the same-named audio file. Importantly, transcriptions should contain no time stamps. If you wish to have a complex set of subfolders within the zip file, as is common for comp-ling corpora like, e.g., NST[4], Autophon will output the finished TextGrids using the same folder structure. An example of one such folder hierarchy is shown in Figure 2

Variationist linguistics:[5] In this mode, you upload pairs of transcription and audio files by the same — one by one or within a zip file. In contrast to the previous mode, transcriptions are longer and include time stamps that delineate the speech at the phrase level. Transcription files may be in Praat `TextGrid` or in ELAN `eaf` and may have multiple speaker tiers. Alternatively, transcription files may be in Excel `xlsx` or a tab-delimited `txt` or `tsv` file.[6] You have the option of uploading a three-column or four-column file, depending on your needs. If the recording has multiple speakers, upload a four-column transcription file whereby column 1 holds the speaker name, column 2 — start time, column 3 — end time, and column 4 — transcription. If the recording has just one speaker, a four-column file is of course fine, but you may also upload a three-column file. Column 1 should hold the start time, column 2 — end time, and column 3 — transcription. Time stamps must be in *seconds formatted as real numbers*. European comma decimals are accepted (e.g., **1,23**) as well as Anglo-American period decimals (e.g., **1.23**). What will not work, however, are time stamps with colons, minutes, or hours (e.g., **00:00:01.23**).

Transcription codecs We have built Autophon so that it accepts transcription files in **most** codecs, and this is a vital feature for its OS-agnostic goal. Accepted codecs include, but are not limited to, UTF-8 Unix, UTF-16 Windows CRLF, Windows ISO Latin 1, and Windows ISO Latin 9. If you encounter errors, please email a sample file to tech support so that we can update our code with that format[7]

Audio codecs We have built Autophon so that it accepts audio files in **most** codecs, and this is a vital feature for its OS-agnostic goal. Accepted codecs include AAC(M4A), AC-3, AIFF, AIFF/24bit, AIFF/32bit, ALAC, FLAC, M4R, MP3, OGG, OPUS, WAV/8bit, WAV/24bit, WAV/32bit, WAV/A-law, WAV/mu-law, and WMA. Autophon will automatically consolidate stereo files to mono, *which may compromise quality due to phase cancellation*[8]. Therefore, you may wish to explicitly select either the left or right channel of your stereo file before aligning. If you encounter errors, please email a sample file to tech support so that we can update our code with that format[9]

Transcription preparation Regardless of what transcription mode you use, transcriptions should contain between one and 20 words. Boundary demarcations should have at least 0.01 seconds of buffer before and after the speech stream. This is illustrated in Figure 3, which shows a five-word phrase with a start boundary 0.03 seconds from the speech and an end boundary 0.25 seconds from the speech. Varying the boundary demarcation in this way is expected, and Autophon handles it well[10]

Select a language Once you upload your files into the aligner, it will suggest a language and language model. You are welcome to change the selection using the dropdown menu.

Task list The task list shows all uploads and includes metrics like file name, upload date, language, tier count, file size, word count, and an inventory of missing words. You can either delete the task and start over, add words to your *custom pronunciations* box (described below), or proceed by clicking *Align*.

---

[4]https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-16/

[5]This also happens to be the field that originally kickstarted forced alignment back in the early 2000s.

[6]The tab-delimited format is similar to the input format that was used for the legacy Penn Forced Aligner and FAVE Align.

[7]In the meantime, a quick fix is to open and resave them in a current version of Praat or ELAN.

[8]For more on phase cancellation, check out https://youtu.be/wY9QokRPJts

[9]In the meantime, a quick fix is to convert the file to WAV using software like FFmpeg or MediaHuman audio converter.

[10]If you have transcriptions of single words or phrases that are segmented at the *exact* start and finish times. Autophon will perform poorly and move those boundaries. This, however, is something we would be interesting in remedying by means of a fifth transcription mode, so kindly reach out to tech support if you have such a project.
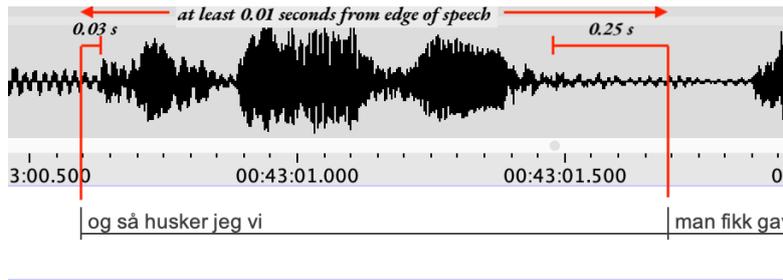
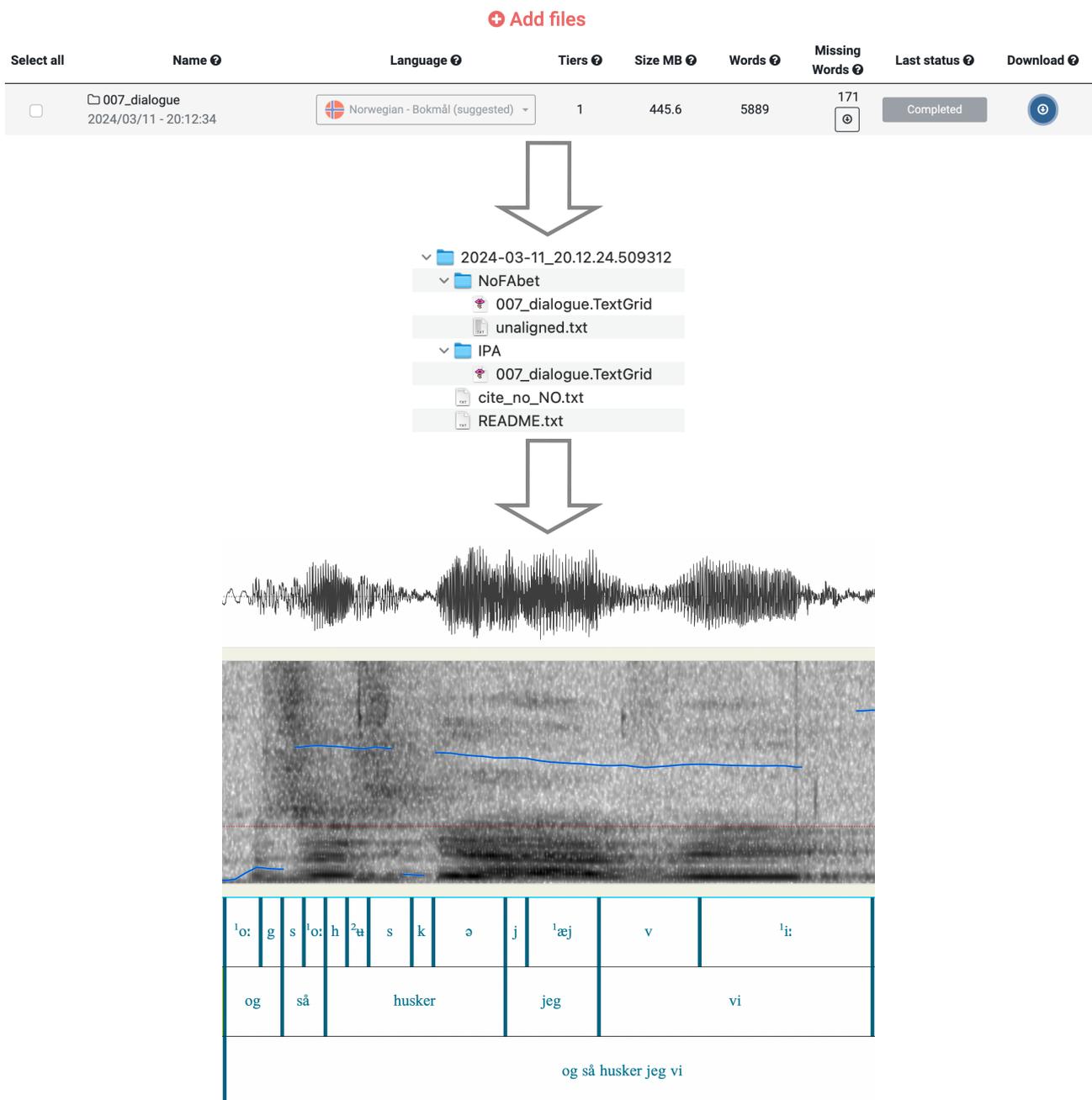*Figure 3: A sample transcription with at least 0.01 seconds of buffer on either end of the speech stream.*

Figure 4: The alignment process, including task list, folder structure, and Praat TextGrid.

Missing words This feature can be understood if you have a basic understanding of how forced alignment works. Forced alignment maps a pre-defined phonemic pronunciation onto the speech stream by means of Deep Neural Networks. These pronunciations are defined by language-specific dictionaries that hold a finite list of words. The *missing words* feature provides a list of words not found in Autophon's dictionary and suggests a corresponding pronunciation. Autophon will simply default to using those suggestions for alignment, but you also can reject a suggestion and enter your own pronunciation. This process is described in the next section.

Your custom pronunciations As described above, forced alignment maps pre-defined phonemic pronunciations onto the speech stream by using language-specific dictionaries that hold a finite list of words. For missing words, Autophon suggests a pronunciation. You may decide that you either (a) do not agree with Autophon's missing words suggestions or that you (b) do not agree with the pronunciations within the language-specific dictionary. In this box you can enter your own pronunciations that will override both.

Pronunciations must be entered using the alphanumeric string specific to the language model at hand — in this case, SweFAbet. Table 1 holds a key for SweFAbet and its respective IPA[11] equivalents. You may type pronunciations directly into the dictionary box or upload them from a **txt** file. You are limited to 1 million characters. Entries must be formatted as **word–space–phoneme–space–phoneme** or **word–tab–phoneme–space–phoneme**. Note that each phoneme must be separated by a space and that every vowel and every diphthong must have a numerical stress attached to it (unstressed vowels take zero). Note also that the lookup cannot be two or more words because that will confuse Autophon and make it treat the second word as a phone.

You may enter more than one pronunciation for the same word by repeating the word on the next line and providing a different pronunciation. Autophon will respond by attempting to find the most suitable pronunciation for that specific speech event. See below for examples of correct versus incorrect entries.

§ *Correct vs. incorrect entries in the "Your Custom Pronunciations" box.*

Correct:

```
dababy     D  AA0  B  EE1  B  II0
dababy     D  B  EJ1  B  II0
da_baby    D  AA0  B  EE1  B  II0
```

Incorrect:

```
dababy     D  AA  B  EE  B  II    (vowel-stress numbering is missing)
dababy     D  AA0  B  EE1  BII0   (phones missing a space between them)
da  baby     D  AA0  B  EE1  B  II0   (two look-ups on a single line)
```

Aligning files Click *Align* to the far right of the upload list to initiate alignment. This will usually just take a few minutes, depending on how many people are using the aligner at that moment.

Downloading the annotations When alignment is finished, your annotations can be downloaded as Praat TextGrids via the downward arrow to the right of the task list. Figure 4 shows an example of this process.

3   How to cite

Any dissemination that makes use of Autophon *Swedish* should cite the below references. We understand that publishers often pressure researchers to slim down bibliographies; however we make our view plain: failure to cite constitutes plagiarism. Refer publishers to this document if you receive pushback.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer software], Version 6.0.36. http://www.praat.org/

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

Young, N. J. (2019b). SweFA 2.0 – Forced Alignment of Swedish, version 2.0. www.autophon.se

Young, N. J., & Anikwe, K. (2024). Autophon – Automatic phonetic annotation of Nordic languages (web application). www.autophon.se

---

[11]International Phonetic Alphabet

4 Phoneme key

Autophon will output two versions of the same TextGrid for every file you align: (1) a TextGrid in the SweFAbet specific to SweFA version 2.0 and (2) a TextGrid in the International Phonetic Alphabet (IPA). The key is located in Table 1. We encourage you to inform us of errors and provide suggestions for changes.

SweFA	IPA	example	SweFA	IPA	example	SweFA	IPA	example	SweFA	IPA	example
Vowels			UH	ɵ	ludd	F	f	fil	RT	ʈ	fart
AA	ɑː	lat	YY	yː	typ	G	g	gas	S	s	sil
AH	a	lass	YH	ʏ	flytta	H	h	hal	SJ	ɧ	sjuk
AE	ɛː	nät				J	j	jag	T	t	tal
AEH	ɛ̞	lätt	**Diphthongs**			JH	d͡ʒ	Jaffar	TH	θ	thriller
EE	eː	leta	AJ	a͡j	tajming	K	k	kål	TJ	ç	tjock
EH	ɛ	lett	AU	a͡u	Gaude	L	l	lös	V	v	vår
II	iː	dis	EJ	ɛ͡j	Facebook	M	m	mil	W	w	wolla
IH	ɪ	disk	EU	ɛ͡u	Europa	N	n	nål	Z	z	guzz
OA	oː	lås	OJ	ɔ͡j	bojkottera	NG	ŋ	ring			
OAH	ɔ	lott				P	p	pil	**Lexical stress and pitch accents**		
OE	øː	söt	**Consonants**			R	r	ris	XX0	x	cirkus
OEH	ø	sött	B	b	bil	RD	ɖ	bord	XX1	^{11}x	cirkus
OO	uː	sot	CH	t͡ʃ	çok	RL	ɭ	Karl	XX2	$_1$x	cirkusdirektör
OH	ʊ	rott	D	d	dal	RN	ɳ	barn	XX3	^{12}x	flytta, cirkusdirektör
UU	ʉː	lus	DH	ð	that's it!	RS	ʂ	fors	XX4	x$_2$	flytta

*Table 1: Phoneme key: SweFAbet, IPA, and lexical examples. The prosodic denotation means that any SweFAbet vowel or diphthong must **always** be followed by the numbers 1 (accent 1), 2 (secondary accent), 3 (accent 2), 4 (post-tonic accent 2), or 0 (unstressed).*
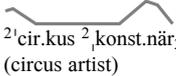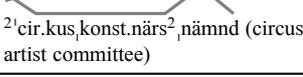
	SMALL accent (usually ω stress)	BIG accent (usually φ head)	Example
Accent 1	H L*	L* H	^{11}cir.kus (circus)
Accent 2	H* L	H*L H	^{21}konst.när$_2$ (artist)
Accent 2 compounds	H* L	H*L *H	^{21}cir.kus $^2{}_1$konst.när$_2$ (circus artist)
			^{21}cir.kus$_1$konst.närs$^2{}_1$nämnd (circus artist committee)

Table 2: Description of pitch accents for Stockholm Swedish as described by (Myrberg and Riad 2015, p. 116): small and big accents 1, 2, and compound accent 2. Examples are shown with orthodox contours. ω stands for word accent, and φ stands for phrase accent (Table from (Young 2019a, p. 39)).

Numerical stress and pitch accent Every SweFAbet vowel is followed by a numerical code that denotes suprasegmental information. XX0 refers to lexically unstressed vowels, XX1 – vowels with acute accent 1, XX2 – vowels with secondary grave accent 2 in compound words, XX3 – vowels with grave accent 2 vowels, and XX4 – unstressed vowels immediately following grave accent 2 vowels (because accent 2 has a delayed peak). The superscript and subscript denotations are adopted from Myrberg and Riad (Myrberg and Riad 2015, p. 116) and are explained more transparently by Table 2, borrowed from Young (Young 2019a, p. 39).

		SweFA 1.0				SweFA 2.0 (current)			
		n boundaries	Median onset difference (ms)	Pct 10 ms or less	Pct 20 ms or less	*n* boundaries	Median onset difference (ms)	Pct 10 ms or less	Pct 20 ms or less
Received Stockholmian	August	1000	13	39	70	1176	10	48	78
	Jan-Axel	1000	16	35	60	1116	12	40	71
	Joseph	1000	13	40	69	1239	11	47	77
Working-class Stockholmian (Ekensnack)	Nils	1000	13	39	66	1240	10	49	78
	Paul	1000	14	33	61	1179	11	44	73
	Per	1000	15	37	62	1200	12	42	72
Stockholm Multiethnolect (Rinkeby Swedish)	Antonio	1000	14	35	63	1241	12	42	72
	Max	1000	12	42	71	1184	11	47	77
	Hayder	1000	13	36	65	1255	12	44	72
	all	9000	13	37	65	10830	11	45	74

Key

n boundaries	number of boundaries tested against the manual gold standard (g.s.)
median onset difference (ms)	median difference between aligner boundaries and manual g.s. boundaries
pct 10 ms or less	percentage of aligner boundaries that fall within 10 milliseconds of manual g.s. boundaries
pct 20 ms or less	percentage of aligner boundaries that fall within 20 milliseconds of manual g.s. boundaries

Table 3: Accuracy metrics for SweFA version 1.0 and the current SweFA version 2.0.

5 Acoustic model and pronunciation dictionary

SweFA version 2.0 for Swedish was trained on the first 36 male speakers[12] in the *Stockholmska Nu!* corpus[13], which consists of spontaneous and read-aloud Stockholm Swedish. The pronunciation dictionary was adapted from The NST Pronunciation Lexicon for Swedish[14].

6 Performance metrics

SweFA version 2.0 meets most of the benchmarks established in the forced-alignment literature. Its accuracy is measured here by comparing alignments of approximately 1000 phonemes in spontaneous speech, each from nine adult male speakers from the Stockholm region. The alignments for the older SweFA version 1.0[15] and the current SweFA version 2.0[16] are compared in Table 3 against manual segmentation.

7 Data security and G.D.P.R.

The files you upload to Autophon are encrypted and sent to a server in Frankfurt, Germany, that is run by *Digital Ocean*. Transcriptions and audio files are deleted immediately after alignment, which significantly reduces the chance of a data breach and keeps our costs low[17]. On the other hand, finished TextGrids are stored in your account for as long as you like. Once, however, you delete them, they will be removed from our server permanently.

---

[12]These are the same speakers analyzed in Young (2019a)

[13]www.stockholmska.nu

[14]https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-22/

[15]Young and McGarrah (2023)

[16]Young (2019b)

[17]We pay Digital Ocean approximately 90 USD per month for 60 GB of space, which means we have thin margins and cannot store much data. This also happens to keep Autophon's carbon footprint relatively low.

If you upload any files and fail to click on *Align*, Autophon will delete them at 3AM Greenwich Mean Time[18].

We recognize our obligations to the European Union General Data Protection Regulation (GDPR), which is why we only collect four types of information from you: name, title, affiliation, and email address. Once you align a file, we permanently delete the audio. Once you delete the file from your task list, we also permanently remove the transcription and documentation of its original name. You may delete your account at any time, at which point we permanently delete your name, title, affiliation, and email address from our server. What we do *permanently* keep, however, is your alphanumeric account ID and the alignment activity linked to that ID — absent of original file names. We keep these records to show funders that Autophon is worth funding.

8 Features and limitations

What Autophon is: Autophon is a frontend web application for the Nordic languages that uses the Montreal Forced Aligner (MFA) [19] as a core component of its backend. The language-specific models and pronunciation dictionaries were constructed by Dr. Nate Young. The most significant pieces of the app's backend were constructed by Kaosi Anikwe who joined the project in early 2023. The language-specific models are trained on various corpora, and the pronunciation dictionaries are usually adaptations of existing dictionaries available online.

The main advantages of using Autophon are:

1. Autophon is a web app, which means it is OS-agnostic.
2. As a web app, it requires no programming knowledge, which expands access to researchers and students.
3. Autophon accepts nearly all types of transcription and sound formats.
4. Autophon has a limitless repertoire of pronunciations by making use of grapheme-to-phoneme algorithms.
5. Autophon has models for Nordic languages, which have typically been neglected by forced alignment tech.

What Autophon is not: Important limitations are:

1. This is no magic bullet. Even with an accurate orthographic transcription, results may not satisfy.
2. Autophon varies in accuracy, and this accuracy depends on the language, speaker, and style.
3. Accuracy metrics are complex projects unto themselves, so they are unavailable for most languages.
4. Autophon will be slower to implement core MFA updates because it consists of layers and layers of code packed around MFA. For example, MFA 2.0 and 3.0 are not part of its backend yet.

9 Budget and funding

Autophon has cost ca. SEK 768 000 (ca. EUR 69 000) to develop and maintain since 2021. It was initially financed with private means by Dr. Nate Young but has since grown in scope with a grant from the Swedish Academy, a grant from the Department of Linguistics and Scandinavian Studies at The University of Oslo, and it has received funding from the European Union's *Horizon 2020 research and innovation programme* under the *Marie Sklodowska-Curie* grant agreement No 892963. Furthermore, The National Library of Norway funded development of the Norwegian Bokmål model[20].

We are actively looking for funders and collaborators who will support Autophon. We are also willing to share authorship with someone who can prepare grant applications and successfully procure funding. Contact us on the support page if you are interested.

Acknowledgements

---

[18]Note that this means that if you are working late at night at, for example, 2.55 AM GMT, your uploaded files may disappear before you manage to align them. Bear this in mind.
[19]McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger (2017)
[20]https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-59/

who helped make the SweFA model possible. Michael McGarrah and Joe Fruehwald offered valuable help for the development of version 1.0.

References

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer software], Version 6.0.36. http://www.praat.org/

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

Myrberg, S., & Riad, T. (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics*, 38(2), 115–147.

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* (tech. rep.). IEEE Signal Processing Society. Piscataway.

Young, N. J. (2019a). *Rhythm in late-modern Stockholm – Social stratification and stylistic variation in the speech of men*. Department of Linguistics, Queen Mary, University of London. ISBN: 978-91-7699-210-4. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-178897

Young, N. J. (2019b). SweFA 2.0 – Forced Alignment of Swedish, version 2.0. www.autophon.se

Young, N. J., & Anikwe, K. (2024). Autophon – Automatic phonetic annotation of Nordic languages (web application). www.autophon.se

Young, N. J., & McGarrah, M. (2023). Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, 46(1), 105–131. \nolinkurl {https://doi.org/10.1017/S033258652100024X}