Autophon user guide
Swedish
Engine: Montreal Forced Aligner 1.0
Model: SweFA 2.0

1 Introducing Autophon and Forced Alignment

Autophon is a free, user-friendly tool for phoneticians that performs *forced alignment* (FA) — the automated process of converting speech recordings and their transcriptions into phonetically time-stamped annotations.

Autophon leverages widely used alignment engines developed by the phonetics community, including:

- *FAVE*[1]
- *faseAlign*[2]
- *Montreal Forced Aligner, version 1.0*[3]
- *Montreal Forced Aligner, version 2.0*[3]

The tool produces time-aligned phonetic annotations compatible with Praat[4], based on two user inputs: (1) a speech recording and (2) its orthographic transcript.

This user guide is specifically for ***Swedish***, using the ***Montreal Forced Aligner 1.0*** engine with the ***SweFA 2.0*** model. Autophon may support additional engine-model combinations for this language; therefore, ensure you are using the best option for your needs.

While many forced aligners exist, they often require command-line usage and are tied to outdated or incompatible operating systems. **Autophon offers a platform-independent, intuitive alternative for phoneticians worldwide**.

2 Using the app

2.1 Aligning files without registering

To align smaller files, go to the main page and click *Add files* at the bottom. A box titled *Transcription Mode: change transcription mode* will appear. Click the heading to choose one of four *Transcription Modes* (see below), then select your files.

2.2 Registering and logging in

To align larger files or access the full suite, click *Sign up* to create a free account. This helps us monitor usage for funders and guard against bots. After signing up, check your email for a verification link. If it doesn't arrive, check your spambox and wait 15 minutes before contacting tech support.

2.3 Cost

Autophon is free of charge.

2.4 Aligning files in a registered account

Once registered and verified, go to the *Aligner* tab and click *Add files*. A box titled *Transcription Mode: change transcription mode* will appear. Click the heading to choose one of four *Transcription Modes*, then select your files.

2.5 Transcription modes

Autophon supports four *Transcription Modes*, named for the fields they're commonly used in: *Experimental Linguistics A*, *Experimental Linguistics B*, *Computational Linguistics*, and *Variationist Linguistics*. Each mode can be selected via the corresponding box in Figure 1, which illustrates expected file structures and links to instructional videos.

Video instructions for each transcription mode can be viewed. In addition, sample templates for each mode are available for **download here**.

[1] FAVE was built by Rosenfelder, Fruehwald, Brickhouse, Evanini, Seyfarth, Gorman, Prichard, and Yuan (2022). It relies on the Hidden Markov Toolkit (S. J. Young, Woodland, and Byrne 1993).

[2] faseAlign was built by Wilbanks (2022). Like FAVE, it relies on the Hidden Markov Toolkit (S. J. Young, Woodland, and Byrne 1993).

[3] The Montreal Forced Aligner was developed by McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger (2017). It uses the Kaldi toolkit (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlicek, Qian, Schwarz, et al. 2011).

[4] Praat is a speech analysis tool developed by Boersma and Weenink (2017).

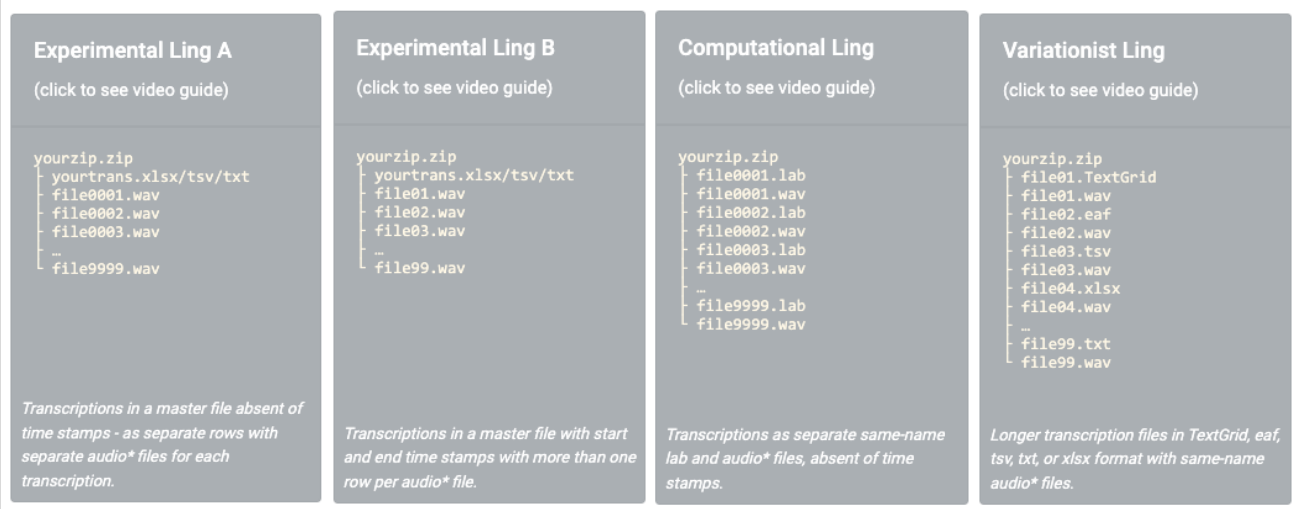

Figure 1: The Transcription Mode selection menu for Autophon.

Figure 2: Autophon outputs finished TextGrids using the same subfolder structure as the uploaded files.

Experimental linguistics A: Upload a two-column spreadsheet (Excel `xlsx`, or tab-delimited `txt`/`tsv`) with audio filenames in column 1 and transcriptions in column 2. No time stamps allowed. This format suits short clips and resembles CommonVoice[5]. Use zip or individual file upload.

Experimental linguistics B: Same structure as A, but with four columns: audio file name, start time, end time, and transcription. Designed for longer recordings requiring segmentation. Time stamps must be in real-number seconds (e.g., `1.23` or `1,23`); no colons or hour-minute markers are permitted (e.g., you may not use something like `00:00:01.23`).

Computational Linguistics: Upload matching audio and `lab` files (containing only the corresponding transcription). Files may be zipped with nested folders—Autophon preserves the hierarchy (Figure 2). No time stamps permitted.

Variationist Linguistics:[6] Upload paired transcription and audio files (individually or zipped). Transcriptions may be in Praat `TextGrid`, ELAN `eaf`, or tabular format (`xlsx`, `txt`, `tsv`). Use either three or four columns:

- Four-column: speaker, start time, end time, transcription
- Three-column: start time, end time, transcription

Time stamps must be real-number seconds (comma or period decimal separators); formats with colons (e.g., `00:00:01.23`) are not supported.

2.6 File formats and codecs

If you encounter errors during upload, it's often due to unsupported file formats or *codecs*. The simplest fix is to re-save your files in a common format using tools like Praat or ELAN.

Transcription file formats: Autophon accepts transcription files in most standard encodings, including `UTF-8` and `UTF-16 (Windows CRLF)`. If you encounter issues, try re-saving the file or email a sample to tech support.

Audio file formats: Autophon supports a wide range of audio formats, including: `WAV`, `FLAC`, `MP3`, and more. Stereo files are not currently accepted. Therefore, convert all audio to mono first.

2.7 Transcription preparation

Regardless of the transcription mode, each entry should contain between one and 20 words. Boundary demarcations must include at least 0.01 seconds of silence before and after the speech. Figure 3 shows a five-word phrase with a 0.03-second pre-boundary and a 0.25-second post-boundary. This sort of variability is expected and handled well by Autophon.[7]

2.8 Select a language

After uploading files into the aligner, Autophon will suggest a language and language model. You may override this suggestion using the dropdown menu.

2.9 Task list

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

2.10 Missing words

To understand this feature, it helps to know that forced alignment maps phonemic pronunciations – defined in language-specific dictionaries – onto the speech stream using statistical models. These dictionaries contain a finite set of words. The *missing words* feature lists items not found in Autophon's dictionary and provides suggested pronunciations. Autophon will use these suggestions by default, but you can reject them and enter your own. The next section explains how.

2.11 Your custom pronunciations

If you disagree with either (a) Autophon's pronunciation suggestions for missing words or (b) the default dictionary entries, you can override them here. Enter your own phonemic transcriptions in this box, which will take precedence over both.

Pronunciations must be entered using the alphanumeric string specific to the language model at hand – in this case, the ***SweFAbet***. Section 4 provides a key that maps the SweFAbet to its IPA[8] equivalents.

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

[6] This field originally drove the development of forced alignment in the early 2000s.

[7] If your transcriptions are segmented with *exact* start and end times, performance may degrade and boundary shifts may occur. If you're working with such data, contact tech support—we are interested in designing a fifth transcription mode for these cases.

[8] International Phonetic Alphabet

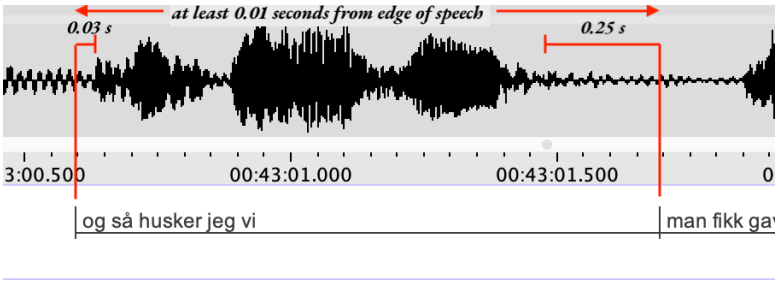

Figure 3: A sample transcription with at least 0.01 seconds of buffer on either end of the speech stream.

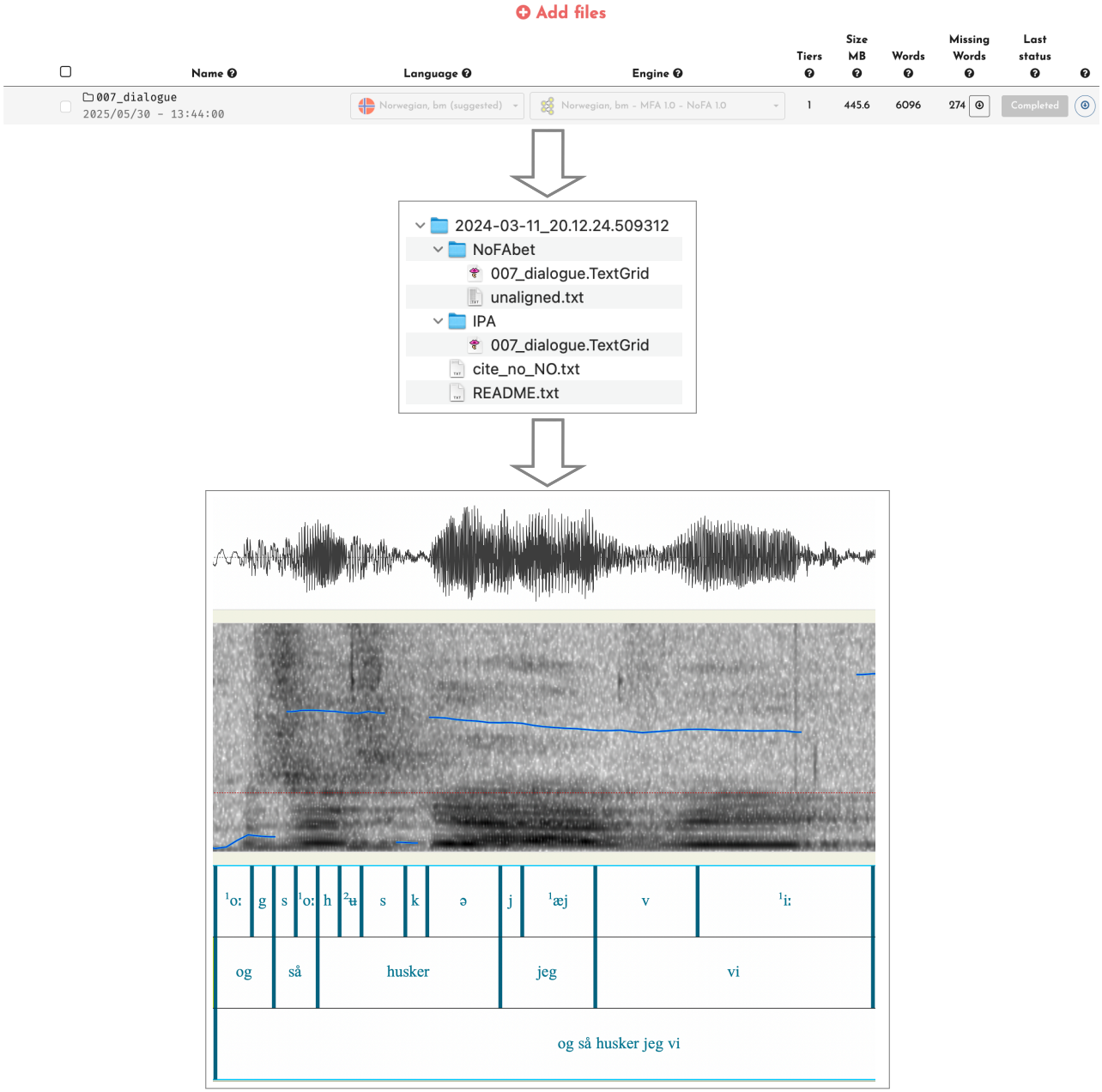

Figure 4: The alignment process, including task list, folder structure, and Praat TextGrid.

You can enter pronunciations directly in the dictionary box or upload them from a `txt` file. The maximum input length is 50 000 characters.

Entries must follow the format:

- `word[space]phoneme[space]phoneme` OR `word[tab]phoneme[space]phoneme`

Each phoneme must be separated by a space, and the lookup may not include two or more words — Autophon will interpret the second word as a phone and produce an error. You may submit multiple pronunciations for the same word by repeating the word on separate lines with different phoneme strings. Autophon will evaluate the best match for each speech instance. Refer to Figure 5 and the examples below.

Stress and/or accent ***must*** accompany every vowel and diphthong with a number. Consult Figure 1 for the specific digits used in this model; refer to Figure 5 to see how these are operationalized.

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)
```

Figure 5: Interface with dictionary entry (left) and phoneme string input (right).

2.12 Aligning files To begin alignment, click *Align* to the far right of the upload list. Alignment typically takes a few minutes, depending on server load.

2.13 Downloading the annotations When alignment is complete, you can download the annotations as Praat TextGrids by clicking the downward arrow beside the task list. See Figure 4 for an illustration.

3 How to cite

Any dissemination or publication that makes use of this Autophon package for ***Swedish***, which uses ***SweFA 2.0*** within ***The Montreal Forced Aligner 1.0*** for its engine, should cite the relevant references listed below. Proper citation is essential: not only to acknowledge the "daisy chain" of technical and academic work underpinning Autophon, but also to reinforce the incentives for sharing tools with the broader community.

While space constraints may tempt you to remove references to software, we strongly encourage prioritizing these citations. If trimming is necessary, consider reducing peripheral citations in the literature review instead.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [softw.], ver.6.0.36. 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. (2019). SweFA 2.0 – Forced Alignment of Swedish, ver. 2.0. www.autophon.org

Young, N. J., & Anikwe, K. H. (2025). Autophon – Automatic phonetic annotation and online forced aligner. www.autophon.org

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 2.0 for The Montreal Forced Aligner 1.0 and (2) a TextGrid in the International Phonetic Alphabet (IPA). The phoneme key is located in Table 1.

SweFA	IPA	example	SweFA	IPA	example	SweFA	IPA	example	SweFA	IPA	example
Vowels			UH	ʉ	*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	ɛ̟	*lätt*	***Diphthongs***			JH	d͡ʒ	*Jaffar*	TH	θ	*thriller*
EE	eː	*leta*	AJ	a͡j	*tajming*	K	k	*kål*	TJ	ɕ	*tjock*
EH	ɛ	*lett*	AU	a͡u	*Gaude*	L	l	*lös*	V	v	*vår*
II	iː	*dis*	EJ	ɛ̟͡j	*Facebook*	M	m	*mil*	W	w	*wolla*
IH	ɪ	*disk*	EU	ɛ̟͡u	*Europa*	N	n	*nål*	Z	z	*guzz*
OA	oː	*lås*	OJ	ɔ̟͡j	*bojkottera*	NG	ŋ	*ring*			
OAH	ɔ	*lott*				P	p	*pil*	***Lexical stress and pitch accents***		
OE	øː	*söt*	***Consonants***			R	r	*ris*	○0	○	*cirkus*
OEH	ø	*sött*	B	b	*bil*	RD	ɖ	*bord*	○1	ˈˈ○	*cirkus*
OO	uː	*sot*	CH	t͡ʃ	*çok*	RL	ɭ	*Karl*	○2	ˌ○	*cirkusdirektör*
OH	ʊ	*rott*	D	d	*dal*	RN	ɳ	*barn*	○3	ˈ2○	*flytta, cirkusdirektör*
UU	ʉː	*lus*	DH	ð	*that's it!*	RS	ʂ	*fors*	○4	○2	*flytta*

Table 1: SweFAbet, IPA, and lexical examples. The denotation for lexical stress and pitch accents 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).*

We encourage you to inform us of errors and provide suggestions for changes.

	SMALL accent (usually ω stress)	BIG accent (usually ϕ head)	Example
Accent 1	H L*	L* H	$^{1}{}^{1}$cir.kus (circus)
Accent 2	H* L	H*L H	$^{2}{}^{1}$konst.när$_{2}$ (artist)
Accent 2 compounds	H* L	H*L *H	$^{2}{}^{1}$cir.kus $^{2}{}_{1}$konst.när$_{2}$ (circus artist)
			$^{2}{}^{1}$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.

Numerical stress and pitch accent Every SweFAbet vowel is followed by a numerical code that denotes suprasegmental information. ○`0` refers to lexically unstressed vowels, ○`1` – vowels with acute accent 1, ○`2` – vowels with secondary grave accent 2 in compound words, ○`3` – vowels with grave accent 2 vowels, and ○`4` – 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 (2015, p. 116) and are explained by Table 2, copied from N. J. 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	m0002	1 000	13	39	70	1 176	10	48	78
	m0008	1 000	16	35	60	1 116	12	40	71
	m0012	1 000	13	40	69	1 239	11	47	77
Working-class Stockholmian (Ekensnack)	m0023	1 000	13	39	66	1 240	10	49	78
	m0024	1 000	14	33	61	1 179	11	44	73
	m0025	1 000	15	37	62	1 200	12	42	72
Stockholm Multiethnolect (Rinkeby Swedish)	m0003	1 000	14	35	63	1 241	12	42	72
	m0020	1 000	12	42	71	1 184	11	47	77
	m0006	1 000	13	36	65	1 255	12	44	72
	all	9 000	13	37	65	10 830	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 2.0.

5 Acoustic model and pronunciation dictionary

This specific Autophon package for ***Swedish*** uses SweFA 2.0 within The Montreal Forced Aligner 1.0, which was trained on the first 36 male speakers[9] in the *Stockholmska Nu!* corpus[10], which consists of spontaneous and read-aloud Stockholm Swedish. The pronunciation dictionary was adapted from The NST Pronunciation Lexicon for Swedish[11].

6 Performance metrics

SweFA 2.0 meets most of the benchmarks established in the forced-alignment literature. Its accuracy is measured here by comparing alignments of approximately 1 000 phonemes in spontaneous speech, each from nine adult male speakers from the Stockholm region. The alignments for the older SweFA version 1.0[12] and the current SweFA 2.0[13] are compared in Table 3 against manual segmentation.

7 Data security and GDPR compliance

Files uploaded to Autophon are encrypted and transmitted to a secure server hosted by *Digital Ocean* within the European Union (Frankfurt and Amsterdam). Transcriptions and audio files are automatically deleted immediately after alignment. This approach enhances privacy while also reducing storage costs. By contrast, finished TextGrids remain available in your account until you choose to delete them. Once deleted, they are permanently removed from our servers.

[9] These are the same speakers analyzed in N. J. Young (2019a)
[10] www.stockholmska.nu
[11] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-22/
[12] N. J. Young and McGarrah (2023)
[13] N. J. Young (2019b)

If you upload files but do not initiate alignment by clicking *Align*, the files will be automatically purged at 3:00 AM GMT[14].

Autophon adheres to the principles of the European Union's General Data Protection Regulation (GDPR). We collect only four pieces of user information: name, title, affiliation, and email address. Once a file is aligned, the corresponding audio is permanently deleted. Deleting a file from your task list also permanently removes the transcription and filename metadata. You may delete your account at any time, which will erase all associated personal data. However, we **do** retain anonymized alignment metadata — such as a randomly assigned alphanumeric user ID and summary usage statistics — to demonstrate the platform's utility to funders.

8 Features and limitations

What Autophon is: Autophon is a web-based application designed to simplify forced alignment workflows and expand access for users with minimal technical background. It is particularly useful for research on under-resourced languages and non-standard varieties, and emphasizes ease of use, format flexibility, and language model diversity. The backend relies on existing forced alignment technologies developed over the past decades, wrapped in a modern frontend that facilitates fast, OS-independent processing.

Key features include:

1. Fully web-based and platform-independent (OS-agnostic).
2. No programming or installation required.
3. Accepts a wide range of transcription and audio formats.
4. Capable of processing low-resource and non-standard language varieties.
5. Supports user-defined pronunciation dictionaries and multiple transcription modes.

What Autophon is not: Important caveats to bear in mind:

1. Alignment quality depends on transcription accuracy, recording quality, and language characteristics.
2. Performance may vary across languages, dialects, and speaking styles.
3. Benchmarking accuracy is ongoing and not available for all models.
4. Core updates to underlying alignment engines may not be immediately reflected, due to the complexity of the Autophon backend.

9 Budget and funding

Autophon costs approximately 25 000 SEK (2 300 EUR) per year to run. Founded by Dr. Nate Young (who is the sole copyright holder), the project has since received support from the University of Helsinki, Linnaeus University, The Swedish Academy, the Department of Linguistics and Scandinavian Studies at the University of Oslo, and the European Union's *Horizon 2020* research and innovation programme under the *Marie Skłodowska-Curie* grant agreement No. 892963. Additional funding for language model development has come from The National Library of Norway[15].

We continue to seek funding and welcome collaboration. If you are experienced in grant writing or interested in supporting the project, please reach out via the support page.

Acknowledgements

Numerous individuals have contributed to Autophon. We especially thank Michael McGarrah for strategic guidance and Kaosi Anikwe for extensive backend and frontend development. Ismail Raji Damilola helped implement a bootstrapping function to expand phoneme inventories. Additional contributions in the early stages came from Nabil Al Nazi, Zamanat Abbas Naqvi, and Santiago Recoba.

We also wish to acknowledge the people who helped make the SweFA model possible. Michael McGarrah and Joe Fruehwald offered valuable help for the development of version 1.0.

[14] Users working near this cutoff time—e.g., at 2:55 AM GMT—should be aware that their files may disappear if alignment is not initiated in time.

[15] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-59/

References

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [softw.], ver.6.0.36. 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.

Rosenfelder, I., Fruehwald, J., Brickhouse, C., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2022). FAVE (Forced Alignment and Vowel Extraction) Program Suite v2.0.0 [Zenodo].

Wilbanks, E. (2022). faseAlign (Version 1.1.14). https://github.com/EricWilbanks/faseAlign

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, ver. 2.0. www.autophon.org

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. doi.org/10.1017/S033258652100024X

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