

Autophon user guide

Danish

Model: DanFA version 4.0

1 What is forced alignment?

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* (McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger 2017). The software 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. For an FA tool to work for a particular language, an acoustic model must be trained and an accompanying pronunciation lexicon must be built that covers every word in the language. FA 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.

2 How to cite

Any study that makes use of Autophon Danish should cite the following sources:

- 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. (2023). Autophon – Automatic phonetic annotation of Nordic languages (web application). www.autophon.se
- Young, N. J. (2024). DanFA 4.0 – Forced Alignment of Danish, version 4.0. www.autophon.se

3 Instructions

Logging in Creating an account and logging in are relatively intuitive. You must verify your email address before you can use Autophon. Wait at least 1 hour for your verification email.

Cost Autophon is free of charge to the public.

Uploading files Files can be uploaded individually, as a group, or in a zip file. Every transcription file must have a corresponding audio file by the same name, e.g., Micke_spontan.TextGrid and Micke_spontan.wav. You can upload/align a maximum of five files, totaling 350 MB, at a time. The following formats are supported:

Transcription formats: Transcription files can be made in Praat (**.TextGrid**) (soon also ELAN **.eaf**) and must only have *one tier each*. Multiple-tier transcriptions will be rejected. If you have a dialogue, extract each tier as a separate file, duplicate the sound file, and give it corresponding names to each extracted tier. TextGrid encoding needs to be either **UTF-8/Unix** or **UTF-16/Windows CRLF**. If you have older Praat files from the mid to early 2000s, they may produce a yellow error box, in which case you should open them on your desktop in a current version of Praat, resave them, and then try uploading them to Autophon again.

Audio formats: Audio files can be **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**, **WMA**. Autophon will automatically consolidate stereo files to mono, which can compromise quality. Therefore, you may wish to convert your audio to mono yourself.

Transcription preparation Transcriptions should be broken down into chunks that are between one and 20 words, and the boundary demarcations should have at least 0.01 seconds of buffer before and after the speech stream. This is illustrated in Figure 1. In natural-speech data, sometimes a researcher will encounter utterances of rapid speech with more than 20 words that have no pauses. In these cases, one must violate one rule in order to appease the other, and it is unclear to us at Autophon what the best approach is. Either force in a break to stick to the 20-word recommendation – violating the 0.01-second buffer – or permit the utterance to go beyond 20 words.

Some users will have transcriptions of single words or phrases that are segmented in Praat down to the exact start and finish time of the speech stream. Autophon will unfortunately perform poorly and not maintain the manual start and finish times. This, however, is something we would be interesting in remedying if a user expresses a need for his/her project.

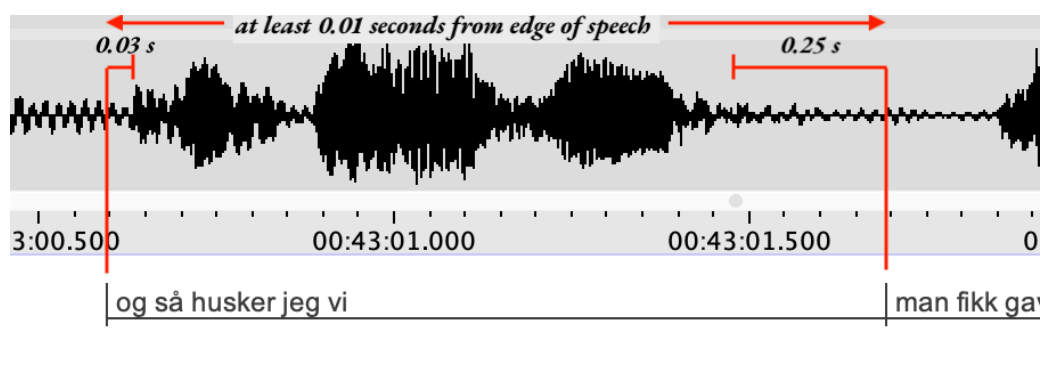


Figure 1: A sample transcription that follows our recommendations: a five-word chunk whereby the start boundary is approximately 0.03 seconds from the speech stream and the end boundary is approximately 0.25 seconds from the speech stream.

Select a language Once the files are uploaded, your files will appear in the topmost upload list where you can proof them before alignment. Autophon will automatically suggest a language for you. You can, of course, change this selection.

Proofing files In the upload list, metrics are provided for your files, including word count, file size, language, and missing words. It is here you can either delete the upload and start over, change the language, add words, or proceed with alignment by clicking “Align!”.

Missing words Alignment is executed with a language-specific model and a language-specific pronunciation dictionary. These pronunciation dictionaries are finite and do not contain every word in the language. Therefore, the “missing words” column outputs any word in your transcription that is not found in Autophon’s dictionary. Specifically, the following process happens: (1) Your transcription is reconciled against Autophon’s dictionary. (2) Missing words are extracted. (3) A grapheme-to-phoneme script that was trained on the original dictionary is run on each missing word to offer suggested pronunciations. (4) These are reported in the “missing words” column, available for download. You then have the option to either accept these suggestions or change them and upload your own changes into the user dictionary (see next section).

Your custom pronunciations You may decide that you either (a) do not agree with Autophon’s suggested pronunciations in the “missing words” column or (b) do not agree with the pronunciations in the back-end dictionary. In this box you can enter your own pronunciations that will override *both*. Entries must be done in the ASCII-bet specific to the language at hand (DanFAbet – Danish, ARPAbet – English, NoFAbet – Norwegian, etc.). The phoneme key can be accessed in Table 1.

Entries can either be made directly into the dictionary box or they can be uploaded from a txt file. You are limited to 1 million characters. Format entries as LOOKUP-SPACE/TAB-PHONEME-SPACE-PHONEME. Note that each phoneme must be separated by a space and that *every vowel 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 phoneme.

If you would like Autophon to optimize and select from more than one pronunciation option, simply enter in multiple pronunciations for the same lookup.

Correct:

```
jabberwocky D J AEEH1 B EH0 W AO2 K IH0
jabberwocky J AEEH1 B EH0 W AO2 K IH0
```

Incorrect:

```
jabberwocky D J AEEH1 B EH W AO2 K IH
jabberwocky D J AEEH BEH WAO2 K IH
jabber wocky J AEEH1 B EH0 W AO2 K IH0
```

Aligning files Click “Align!” to the far right of the upload list to initiate alignment. This can take between a few minutes and an hour, depending on the amount of users accessing the aligner at that moment. You may have to log out and then log back in to see the outputted files.

Downloading the annotations Once alignment is finished, your annotations will be available in the bottom-side download list. You may have to log out and then log back in to see the outputted files. Click on the item to download the folder. The annotations will be in **TextGrid** format and are only readable in Praat (Boersma and Weenink 2017).

DanFA	IPA	ex.	DanFA	IPA	ex.	DanFA	IPA	ex.	DanFA	IPA	ex.
Full vowels			OEH	ø	<i>synder</i>	MQ	m ₁	<i>dame > dam</i>	R	ʁ	<i>rose</i>
AA	ɑ:	<i>far</i>	OH	o	<i>foto</i>	NQ	n ₁	<i>båne > bån</i>	S	s	<i>si</i>
AE	ɛ:	<i>kæde</i>	OO	o:	<i>ro</i>	NGQ	ŋ	<i>konge > kon</i>	SJ	ɛ	<i>sjæl</i>
AEE	æ:	<i>mast</i>	UH	u	<i>guld</i>	Consonants			T	t	<i>Tid, tjak</i>
AEEH	æ	<i>mat</i>	UU	u:	<i>bus</i>	B	b	<i>bil</i>	TH	θ	<i>Thanksgiving</i>
AEH	ɛ	<i>mæt</i>	YH	y	<i>nyt</i>	D	d	<i>dag</i>	V	v	<i>vase</i>
AH	ɑ	<i>takke</i>	YY	y:	<i>ny</i>	DX	ð	<i>bad</i>	W	ʊ	<i>hav</i>
AO	ɔ:	<i>morse</i>	AXH	ʌ	<i>råt</i>	F	f	<i>fin</i>	Z	z	<i>Benz</i>
AOH	ɔ	<i>vor</i>	UXH	ə	<i>midte</i>	G	g	<i>gul</i>	ZH	ʒ	<i>jub</i>
EE	e:	<i>vene</i>	Diphthongs			H	h	<i>bæs</i>	Lexical stress and glottal stød		
EH	e	<i>midt</i>	-J	-j	<i>nej</i>	J	j	<i>ja</i>	AAAs1	'ɑ:ʔ	<i>b̥arn</i>
IH	i	<i>mit</i>	-R	-r̥	<i>jer</i>	K	k	<i>kost</i>	AA1	'ɑ:	<i>b̥arnevogn</i>
II	i:	<i>vin</i>	-W	-v̥	<i>bav</i>	L	l	<i>land</i>	AAAs2	'ɑ:ʔ	<i>sp̥ædb̥arn</i>
OA	ɔ:	<i>rå</i>	Schwa assimilation			M	m	<i>man</i>	AA2	'ɑ:	<i>sp̥ildv̥arme</i>
OAH	ɔ	<i>rust</i>	DXQ	ð̥	<i>bade > bað̥</i>	N	n	<i>nord</i>	AA0	ɑ:	<i>sp̥ædb̥arnspleje</i>
OE	ø:	<i>føre</i>	LQ	l̥	<i>tale > tal̥</i>	NG	ŋ	<i>eng</i>			
OEE	œ:, œ:	<i>fyrre</i>				P	p	<i>pil</i>			
OEEH	œ, œ	<i>sønner</i>									

Table 1: Phoneme key: DanFabet, IPA, and lexical examples

4 Phoneme key

Autophon will output two versions of the same TextGrid for every file you align: (1) in DanFabet and (2) in IPA. DanFabet is the ASCII-based phoneme coding that is specific to the Danish language and resembles CMU’s ARPAbet¹. The key is located in Table 1.

Every DanFabet vowel is followed by a numerical code that denotes suprasegmental information. XX0 refers to lexically unstressed vowels; XX1 – primary lexical stress; XXs1 – primary lexical stress with stød; XX2 – secondary lexical stress; XXs2 – secondary lexical stress with stød.

5 Performance and metrics

The current DanFA version 4.0 (Young 2021) uses the same model as DanFA 3.0, albeit with an improved pronunciation dictionary and phoneme inventory that reflects more updated understandings of Danish phonology. The language model was trained on spontaneous and read-aloud Copenhagen Danish from the DanPASS corpus (Grønnum 2006), read-aloud speech from the NST Danish ASR database (<https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-55/>), and the Nordic Dialect Corpus (Johannessen, Priestley, Hagen, Nøklestad, and Lynum 2012). This upgrade is now quite accurate and exceeds 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 male teenage speakers from the Copenhagen region. The alignments are compared with manual segmentation in Table 2.

6 Data security

Everything you upload is encrypted and sent to a server in Frankfurt, Germany, that is run by *Digital Ocean*. Transcriptions and sound files are deleted immediately after alignment, which significantly reduces the chance of a data breach. 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.

7 Features and limitations

What Autophon is

¹<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

		DanFA 1.0				DanFA 2.0				DanFA 3.0/4.0 (current)			
		<i>n</i> boundaries	Median onset difference (ms)			<i>n</i> boundaries	Median onset difference (ms)			<i>n</i> boundaries	Median onset difference (ms)		
			Pct 10 ms or less	Pct 20 ms or less		Pct 10 ms or less	Pct 20 ms or less		Pct 10 ms or less	Pct 20 ms or less		Pct 10 ms or less	Pct 20 ms or less
Received Copenhagen	Speaker 1	1128	14	34	62	1163	12	44	67	1163	10	47	72
	Speaker 2	1031	16	32	57	1265	10	49	70	1265	10	51	77
	Speaker 3	1062	17	34	55	1474	9	53	73	1474	10	50	76
Working-class Copenhagen	Speaker 1	639	17	30	55	638	15	36	59	638	11	44	69
	Speaker 2	804	13	37	63	806	14	41	63	806	10	45	73
	Speaker 3	740	18	30	53	750	11	47	66	750	10	46	75
Copenhagen Multiethnolect	Speaker 1	845	19	28	51	872	2	59	76	872	10	46	68
	Speaker 2	743	17	32	55	726	18	35	54	726	13	41	61
	Speaker 3	1011	17	32	54	1063	10	48	66	1063	10	43	69
<i>all</i>		8003	16	33	57	8757	11	47	67	8757	10	47	72

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

Table 2: Accuracy metrics for DanFA version 1.0 (Young 2017), DanFA version 2.0 (Young 2021), DanFA version 3.0 (Young 2021), and the current DanFA version 4.0 (Young 2024)

Autophon is a frontend web application for the Nordic languages that uses the Montreal Forced Aligner (MFA) (McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger 2017) as a core component of its backend. The language-specific models and pronunciation dictionaries were constructed by Nate Young. 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, Autophon also requires less computational knowledge, which expands access to more researchers and students.
3. Autophon takes nearly all types of transcription and sound formats. It runs your transcriptions and sound files through a converter to ensure that they can be aligned with the MFA backend.
4. Autophon offers an easy way to add the pronunciation of missing words by integrating grapheme-to-phoneme algorithms into the user interface.
5. Autophon allows the user to have access to the latest language training models without needing to constantly check for updates.

What Autophon is *not*

Important limitations are:

1. Autophon is no magic bullet. You need a highly accurate orthographic transcription for it to work. And even then, you may not be satisfied with the results.
2. Autophon varies in its accuracy, and this accuracy depends on the language. Accuracy metrics are provided above.
3. Autophon will be slower to implement core MFA updates because it consists of layers and layers of code packed around MFA.

Acknowledgements

Numerous individuals helped make DanFA possible. Michael McGarrah and Joe Fruehwald assisted me with developing version 1.0. Kaosi Anikwe has been an invaluable backend and frontend developer. Ismail Raji Damilola, with his amazing AI know-how, has been invaluable for developing versions 3.0 and 4.0. Nicolai Pharaos and Bjørn Falkencrone worked patiently to offer feedback

on how to overhaul the pronunciation dictionary from its original roughshod form, thereby enabling the release of version 4.0. Nina Grønnum was kind enough to make her DanPASS recordings accessible, upon which versions 2.0 and 3.0 were trained. Nicolai Pharao, Christophe Zerkitsky Vies, and Rebecca Borg assisted me with validation and testing. I also have the following programmers to thank, all of whom worked tirelessly to build a robust backend and UX: Nabil Al Nazi, Zamanat Abbas Naqvi, and Santiago Recoba.

References

- Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer software], Version 6.0.36. <http://www.praat.org/>
- Grønnum, N. (2006). DanPASS - A Danish Phonetically Annotated Spontaneous Speech Corpus. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. http://www.lrec-conf.org/proceedings/lrec2006/pdf/4_pdf.pdf
- Johannessen, J. B., Priestley, J., Hagen, K., Nøklestad, A., & Lynum, A. (2012). The Nordic Dialect Corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3387–3391. http://www.lrec-conf.org/proceedings/lrec2012/pdf/773_Paper.pdf
- 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. (2017). DanFA 1.0 – Forced Alignment of Danish, version 1.0. <https://github.com/mcgarran/LG-FAVE>
- Young, N. J. (2021). DanFA 2.0 – Forced Alignment of Danish, version 2.0.
- Young, N. J. (2023). Autophon – Automatic phonetic annotation of Nordic languages (web application). www.autophon.se
- Young, N. J. (2024). DanFA 4.0 – Forced Alignment of Danish, version 4.0. www.autophon.se