Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS

Jan Strunk, Florian Schiel, Frank Seifart

University of Amsterdam, Bavarian Archive for Speech Signals (BAS), University of Amsterdam
j.strunk@uva.nl, schiel@bas.uni-muenchen.de, F.C.Seifart@uva.nl

Abstract

Language documentation projects supported by recent funding intiatives have created a large number of multimedia corpora of typologically diverse languages. Most of these corpora provide a manual alignment of transcription and audio data at the level of larger units, such as sentences or intonation units. Their usefulness both for corpus-linguistic and psycholinguistic research and for the development of tools and teaching materials could, however, be increased by achieving a more fine-grained alignment of transcription and audio at the word or even phoneme level. Since most language documentation corpora contain data on small languages, there usually do not exist any speech recognizers or acoustic models specifically trained on these languages. We therefore investigate the feasibility of untrained forced alignment for such corpora. We report on an evaluation of the tool (Web)MAUS (Kisler et al., 2012) on several language documentation corpora and discuss practical issues in the application of forced alignment. Our evaluation shows that (Web)MAUS with its existing acoustic models combined with simple grapheme-to-phoneme conversion can be successfully used for word-level forced alignment of a diverse set of languages without additional training, especially if a manual prealignment of larger annotation units is already avaible.

Keywords: word times, forced alignment, language documentation corpora

1. Introduction

1.1. Language Documentation Corpora

Recent years have seen major global efforts in the paradigm of language documentation, which aims at providing a comprehensive multimedia record of the linguistic practices of speech communities (Himmelmann, 1998, p. 166), especially of those speaking less-studied and endangered languages. Supported by large funding initiatives, such as DoBeS (Dokumentation bedrohter Sprachen) by the Volkswagen foundation,[1] the US National Science Foundation and National Endowment for the Humanities' Documenting Endangered Languages Program, or the Hans Rausing Endangered Languages Project,[2] language documentation projects have created rich multimedia corpora of many different and typologically diverse languages. These language documentation corpora usually comprise audio and video recordings of spoken language, most of which have also been transcribed in a practical orthography, and often also linguistic annotations such as interlinear glossing of words and morphemes. In most recent language documentation projects, the transcription has been manually aligned with the audio and/or video material at the level of relatively large units, such as utterances, sentences, intonation units, or paragraphs, while legacy data may consist of recordings and transcriptions separately without any kind of alignment between them. These spoken-language corpora represent unique and novel resources for typological, corpus linguistic, psycholinguistic, and sociolinguistic research. Moreover, they are very valuable resources for the creation of learning tools and other language resources. Their usefulness, however, could be greatly increased with a more fine-grained temporal alignment of transcription and audio/video data at the level of words, syllables, or individual phones. This would enable novel corpus-based research which fully embodies that "language is a temporal phenomenon, a process that flows through time" (Chafe, 2002, p. 256). Manual alignment of transcription and audio at the word or even phone level would require immense amounts of time and manpower. We therefore investigate here the possibility of using automatic forced alignment in the form of the tool WebMAUS (Kisler et al., 2012) in order to obtain a word-level alignment of transcription and audio. Without any additional training on the specific languages of our corpora, WebMAUS produces quite encouraging alignment results that we consider to be of sufficiently high quality for use in actual linguistic analyses.

[1] http://dobes.mpi.nl/

[2] http://www.hrelp.org/

1.2. Practical Use Case: A DoBeS Comparative Corpus Analysis Project

The comparative corpus analysis project "The relative frequency of nouns, pronouns, and verbs cross-linguistically" (Seifart et al., 2010; Seifart, 2011), funded by the DoBeS initiative of the Volkswagen foundation, examines the ratio of nouns to verbs in corpora of several typologically diverse languages; cf. Table 1. In addition to studying variations in the noun-to-verb ratio (NTVR) from a typological perspective, correlating it with typological characteristics of a language such as the extensiveness of argument indexing and its basic word order, and from a sociological and stylistic perspective, taking speaker characteristics and text genres into account, this project also examines the development of the NTVR in real time as narrative texts and conversations unfold. Pilot studies reported in Seifart et al. (2010) and Seifart (2011) have revealed a characteristic temporal pattern with a relatively high noun-to-verb ratio at the beginning of texts and subsequent sinusoidal alternations as the narrative unfolds. We also investigate possible correlations between the NTVR and processing ease as reflected in speech rate. The existing manual segmentation and align-

	Language			Corpus	
Language	Affiliation	Region	Speakers	Words	Source
Baure	Arawakan	West Amazon	55	27,907	Swintha Danielsen et al.
Bora	Boran	North-West Amazon	1,500	29,539	Frank Seifart
Chintang	Sino-Tibetan	Himalaya	4,500	37,050	Balthasar Bickel et al.
Even	Tungusic	Siberia	300	36,665	Brigitte Pakendorf
Hoocąk	Siouan	USA	200	23,503	Iren Hartmann et al.
N|uu	Tuu (South. Khoisan)	South Africa	6	32,126	Tom Güldemann et al.
Sakha	Turkic	Siberia	360,000	30,850	Brigitte Pakendorf
Sri Lanka Malay	Austronesian	Sri Lanka	45,000	13,044	Sebastian Nordhoff
Texistepec Popoluca	Mixe-Zoquean	Mexico	100	24,674	Søren Wichmann

Table 1: Languages investigated in the noun-to-verb ratio project

ment of texts into larger annotation units is not adequate for this purpose, because, firstly, the criteria for establishing these units are not comparable between the different language documentation corpora, and, secondly, the units are also relatively large and alignment thus coarse. Therefore, we decided to use words as a less variable and more fine-grained unit in our time-series analyses of the noun-to-verb ratio and to link them directly to the timeline using automatic forced alignment methods, possibly combined with a manual correction stage.

2. Automatic Forced Alignment

2.1. Previous Work

Untrained forced alignment of individual phones has already been explored with encouraging results for isolated words in a Mixtec corpus by DiCanio et al. (2013). Unlike this study, we present and evaluate a method to obtain accurate word start and end times for words in the context of complete spoken texts (that are up to one hour long) based on the WebMAUS service of the Bavarian Archive for Speech Signals (BAS) (Kisler et al., 2012).

2.2. The WebMAUS Automatic Alignment System

The Munich AUtomatic Segmentation system (MAUS) and the corresponding CLARIN web service WebMAUS combine simple forced alignment based on Hidden Markov Modeling (HMM) with optional additional statistical modeling of possible pronunciation variants for several languages. The aligner has the task to find the best partitioning of the speech signal given a statistical pronunciation model and a set of pre-trained acoustical models (HMM) for each phoneme class of a language. Forced alignment works very well granted that the signal is of moderate good quality and the truly spoken phones are known a priori, that is, the input transcription is relatively accurate. (Web)MAUS extends the basic HMM aligner concept by modeling a statistical space of possible pronunciation variants for a given orthographic input (Schiel, 1999; Schiel, 2004). For known languages, the hypotheses space is calculated for each individual text input based on a machine-learned statistical expert system of pronunciation (Schiel et al., 2011). Combined with HMM technology, the MAUS can thus not only find the best segmentation but at the same time the most likely sequence of truly spoken phones in the speech signal. On a subset of spontaneous German speech in the Verbmobil corpus (Burger et al., 2000) the MAUS technique yielded about 97% of the average interlabeler agreement of three trained phoneticians working on the same task (Kipp et al., 1996). MAUS is implemented as a system of UNIX script files and C++ binaries that can be run on Linux and Windows platforms. It requires as input the speech signal and some form of either orthographic or phonological transcript of the spoken utterance. The result is stored in either BAS Partitur Format BPF (Schiel et al., 1998), praat TextGrid or Emu (Bombien et al., 2006) compatible annotation format files. MAUS currently (version 2.68) supports 11 languages: German, Polish, Portuguese, English, Australian English, New Zealand English, Hungarian, Italian, Estonian, Spanish, Dutch, and a special language independent mode called 'sampa', that allows the segmentation of arbitray languages encodedable in SAM-PA. A vast number of options allow the user to control the alignment process as well as the form of output formatting and the statistical modeling of the pronunciation variation. The MAUS freeware package can be downloaded from the Bavarian Archive for Speech Signals;[3] a web interface to a server based implementation called WebMAUS is also available[4] (Kisler et al., 2012).

2.3. Specific Forced Alignment Procedure

Since we are dealing with small endangered languages in our comparative corpus analysis project, for which WebMAUS has no pretrained models of pronunciation variation, we use it in a simple mode based only on HMM forced alignment without modeling possible pronunciation variants. For all the results reported throughout this paper, we used the special 'sampa' mode of WebMAUS, which combines the acoustic models of languages that MAUS currently supports, in order to have as large a phonetic inventory as possible for the alignment of our diverse set of languages. We thus do not carry out any training or adaptation of WebMAUS to the specific languages in our corpus.
A schematic of the workflow we use for forced alignment in our project is given in Figure 1. The first step required in forced alignment is grapheme-to-phoneme conversion, that is, the conversion of the original orthography of the

[3] http://www.bas.uni-muenchen.de/forschung/Bas/software/
[4] http://clarin.phonetik.uni-muenchen.de/BASWebServices/

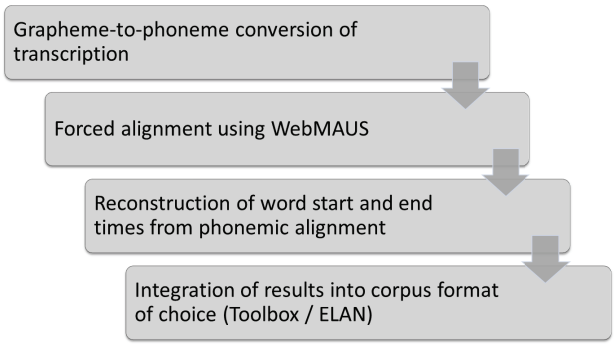

Figure 1: Workflow of forced alignment based on WebMAUS

Orthography	MAUS SAM-PA inventory	IPA
aka	a k a	aka
muhdú	m M Q d M	mɯʔdɯ́
múúne	m M M n E	mɯ́ːnɛ
iiñúhejúne	i i J M Q E h M n E	iiɲɯ́ʔɛhɯ́nɛ
péétsó	p E E ts o	pɛ́ɛ́tsó
uubálle	M M b a tS E	ɯːbádʒɛ
cána	k a n a	kána
oke	o k E	okɛ
duubálle	d M M b a tS E	dɯːbádʒɛ
uúh	M M Q	ɯɯ́ʔ
diibyéváa	d i i b j E B a a	diːbiɛ́váa
iiñúhejúne	i i J M Q E h M n E	iːɲɯ́ʔɛhɯ́nɛ
péétsó	p E E ts o	pɛɛːtsó

Figure 2: Grapheme-to-phoneme conversion for a Bora example

transcriptions into a phonemic transcription suitable for the aligner used. Since we are dealing with languages for which WebMAUS does not provide ready-to-use phoneme-to-grapheme conversion modules, we perform this step ourselves using very simple transducers. Specifically, we use a short ordered sequence of simple search-and-replace rules to remove capitalization, delete punctuation marks, and to convert every grapheme in the orthographic transcription of the respective language into the SAM-PA inventory provided by WebMAUS' special 'sampa' mode. This inventory includes a large part of the IPA inventory but no click sounds, for example, which occur in one of our corpora, namely the N|uu corpus, or specific models for retroflex stops required for the phonology of Sri Lanka Malay (SLM) (cf. section 3.2.). The grapheme-to-phoneme conversion is thus an approximation limited by the set of acoustic models provided by WebMAUS and based on intutions about acoustic and articulatory similarity. It is carried out using a Python script that also converts the input corpus session files, usually Toolbox files,[5] typically used in language documentation projects, or ELAN files (Wittenburg et al., 2006),[6] to the input BAS Partitur Format BPF of WebMAUS (Schiel et al., 1998). An example of an automatic conversion from Bora practical orthography into the SAM-PA representation used by WebMAUS is given in figure 2. The rightmost column also provides an IPA transcription for comparison.

In the second step, the converted transcription and the corresponding audio file are then uploaded to the WebMAUS web service.[7] We use the 'General MAUS' variant of WebMAUS with the SAM-PA language, no modelling of pronunciation variants (option 'Canonly' set to `true`) and BPF output format (option 'mau-append'). Depending on whether we would like to constrain the alignment process on the basis of a pre-existing manual alignment at the level of larger units, such as paragraphs, sentences, or intonation units, or not (see below), we set the option 'Usetrn' to `true` for constrained alignment and to `false` for unconstrained alignment. The other avaible options are left at their default values.

[5] `http://www-01.sil.org/computing/toolbox/`

[6] `http://tla.mpi.nl/tools/tla-tools/elan/`

[7] WebMAUS also provides a way to automate uploading and downloading using cURL (`http://curl.haxx.se/`).

The result output by WebMAUS, also in BAS Partitur Format, with alignment on the phoneme-level is then converted back to Toolbox or ELAN format using another set of Python scripts. Since we are currently interested in the alignment of whole words, in the third step, we reconstruct the start and end times of words from the alignment of individual phones provided by WebMAUS and finally integrate these word times directly into the session files in the case of ELAN files or store them in word-level tiers in Toolbox files for later statistical analysis.

3. Evaluation

In this section, we report on three studies in which we have evaluated the feasibility of using WebMAUS for the untrained forced alignment of language documentation corpora. The first study discussed in section 3.1. compares the performance of WebMAUS on eight small texts from five different languages to differences between two human aligners. Section 3.2. contains a more practical evaluation of WebMAUS as we have actually used it to align transcriptions and audio for the corpora in our NTVR project. Last but not least, we test WebMAUS on a larger corpus of Hoocąk recordings which were manually aligned at the word-level independently of our project (section 3.3.).

Throughout this section, we compare what we call 'unconstrained' and 'constrained' forced alignment. Unconstrained alignment only provides WebMAUS with the audio data and the transcription itself and no additional information about where to look for particular words. WebMAUS therefore tries to align the given sequence of words from the transcription in strict linear order from the start of the audio file to its end. Unconstrained alignment could, for example, be used to align legacy recordings with their separately stored transcriptions. This mode of alignment can be problematic for transcriptions with gaps of non-transcribed stretches of the audio recording or for transcriptions of dialogs with overlapping turns (since words from overlapping turns are normally not interleaved in the correct order in tools such as Toolbox) (cf. section 4.). In the case of recently compiled language documentation cor-

pora, the transcription is often already aligned with the audio recording as part of the transcription process (for example, in ELAN), albeit in larger annotation units, such as utterances, intonation units, sentences, or paragraphs. Constrained alignment provides WebMAUS with the start and end times of such pre-existing manually aligned annotation units and thus with information about the time stretches in which to search for particular words. This alignment mode is suitable for most recent corpora of spoken language that already contain explicit links between parts of the transcription and stretches in the audio recording. Since WebMAUS allows for a pre-segmentation with overlapping chunks, constrained alignment can also be used to align words when the contributions of several speakers overlap.

3.1. Evaluation on Small Test Corpora in Comparison to Differences between Two Human Aligners

Before we decided to use WebMAUS for the forced alignment of our entire NTVR project corpus (cf. section 3.2.), we carried out a small evaluation study[8] in order to test its performance with data from several languages and with texts that we deemed relatively easy as well as with texts that we regarded as relatively hard to align. For three languages, namely, Baure, Bora, and Even, we tested WebMAUS with one 'easy' text and one 'hard' text each. The 'easy' texts are simple monological narrative texts with a reasonably good audio quality and few background noises; the 'hard' texts contain contributions from several speakers, including overlaps, as well as background noises (such as traffic or animals) and, in the case of Bora, even several people cheering and screaming. For the remaining languages, German and Sri Lanka Malay, we only tested one 'easy' narrative text.[9] The words in these eight texts were also manually aligned independently by two human aligners, once by the first author and once by a student assistant. In order to speed up the alignment process, the human aligners were provided with the pre-aligned larger annotation units and exactly aligned word starts and endings within these larger units.

Table 2 compares the results of automatic alignment at the word-level using WebMAUS to manually aligned word start and end times for the eight test sessions. Results are given as mean, median, and maximum (unsigned) time differences measured in milliseconds (ms). Both word start and end times are included in these figures. The German results are included for comparison. They were obtained without using the grapheme-to-phoneme conversion for German provided by WebMAUS. The last three columns of Table 2 also provide mean, median, and maximum time differences between two human aligners for comparison.

Looking at the mean time differences in table 2, especially those for unconstrained alignment, one could initially get the impression that the alignment quality obtained by using WebMAUS for untrained forced alignment on 'unknown' languages is less than ideal. Unconstrained alignment yields word start and end times that are as much as as 6.5 seconds (Bora hard) or even 15.3 seconds (Baure easy) off on average. However, except for the hard Bora session, the median, as a measure of central tendency that is more robust to extreme outliers, shows that most automatically aligned word times are reasonably close but that forced alignment, especially in the unconstrained alignment mode, sometimes goes completely astray for parts of a transcription and that these extreme outliers cause the relatively high mean time differences for unconstrained alignment; compare the maximal time differences of almost two minutes for the easy Baure session and maximal differences between 3 and 25 seconds for the other sessions. The misalignments for the easy Baure session and the hard Bora session can be explained by the fact that, in both sessions, parts of the audio were left untranscribed: In the case of the hard Bora session, there are gaps in the middle of the transcription, whereas in the case of the easy Baure session, the beginning and ending of the audio file were not completely transcribed because they contained introduction and farewell etc. in Spanish. The boxplots in Figure 3 nicely illustrate the observation that most word times obtained by unconstrained or constrained forced alignment are quite accurate but that unconstrained alignment can be completely off for some stretches of a recording (usually due to untranscribed parts of the audio file and/or overlapping utterances) (note the high number of outliers in the leftmost boxplot), whereas constrained alignment (the central panel in Figure 3) does not exhibit as many large errors because the aligner can only get confused within the time stretch of one larger pre-aligned annotation unit. Human aligners are of course even less prone to get completely confused by missing parts in the transcription or other problems, as shown by the lack of outliers in the rightmost boxplot.

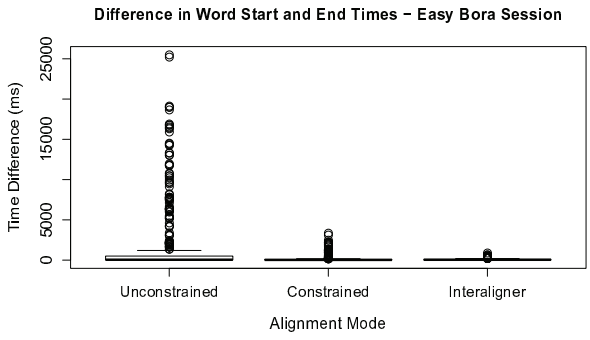

Figure 3: Word time differences for the easy Bora session

A statistical analysis of the results in Table 2 shows that, for all test sessions, the time differences between the human gold standard and the results of forced alignment are significantly smaller for constrained alignment than for unconstrained alignment, except on the Sri Lanka Malay session where there is no difference between unconstrained and constrained alignment.[10] For all test session, human in-

[8] Initial results of this evaluation study were already presented at the 2013 DoBeS conference in Hannover (Strunk, 2013).

[9] The German text was created within the project AUVIS (Audiovisual Data-Mining for Event Segmentation in Multimodal Speech Data) at the University of Cologne.

[10] I report the result of unparametric Wilcoxon tests here: Baure easy ($W = 746237$, $p < 0.001$), Baure hard ($W = 1159270$,

		Unconstrained Alignment			Constrained Alignment			Interaligner Differences		
Test Session	Words	mean	median	max	mean	median	max	mean	median	max
Baure (easy)	502	15,312 ms	300 ms	116,810 ms	160 ms	101 ms	1,030 ms	115 ms	0 ms	1,063 ms
Baure (hard)	689	1,100 ms	139 ms	18,361 ms	204 ms	60 ms	3,214 ms	86 ms	48 ms	1,932 ms
Bora (easy)	289	1,455 ms	50 ms	25,496 ms	148 ms	30 ms	3,338 ms	76 ms	45 ms	910 ms
Bora (hard)	108	6,460 ms	8,020 ms	12,995 ms	290 ms	160 ms	1,485 ms	162 ms	66 ms	1,398 ms
Even (easy)	405	696 ms	37 ms	12,833 ms	196 ms	34 ms	2,272 ms	57 ms	26 ms	745 ms
Even (hard)	236	612 ms	183 ms	5,820 ms	248 ms	63 ms	2,589 ms	59 ms	30 ms	1,193 ms
German (easy)	467	131 ms	31 ms	3,172 ms	42 ms	32 ms	770 ms	25 ms	17 ms	170 ms
SLM (easy)	204	297 ms	38 ms	6,180 ms	207 ms	41 ms	6,127 ms	104 ms	41 ms	2,607 ms

Table 2: Time differences between automatic alignment using WebMAUS and human aligners (word start and end times)

terannotator differences are also significantly smaller than the differences between one human annotator and the results of constrained alignment, except again for the Sri Lanka Malay session where the test fails to reach significance.[11] Even though WebMAUS makes use of acoustic models trained on German speech data, rather than speech data from Baure, Bora, Even, or Sri Lanka Malay, the median time differences for these other languages seem to be roughly comparable to the results for German. Comparing constrained alignment on the easy German session to all other easy sessions, however, shows that the forced alignment results on the German session are still significantly closer to the human alignment gold standard than the other languages.[12] This is probably due both to the acoustic models of WebMAUS as well as the fact that the transcription of the German session is exceptionally detailed and complete. Interestingly, the intuitive distinction between easy and hard sessions (for Baure, Bora, and Even) is only reflected in the constrained automatic alignment results, whereas, for unconstrained alignment, the accuracy and completeness of the transcription seems to be more relevant to successful alignment than the acoustic difficulty of a session (background noises, overlaps, etc.).[13] The distinction also does not seem to be very relevant to human aligners.[14]

3.2. Evaluation on the Entire Corpus of the Noun-to-Verb Ratio Project

Based on the results of our pilot evaluation reported in the previous section, we decided to go ahead and use WebMAUS to produce a word-level alignment between audio and transcription for all texts in our entire NTVR project corpus. Because most of our subcorpora already contained a pre-existing manual alignment at the level of larger annotation units and because unconstrained alignment can easily be led astray by incomplete transcriptions and speaker overlaps, as we have seen in the previous section, we used WebMAUS in constrained alignment mode whenever possible. In the case of the Texistepec Popoluca corpus, for which there existed no previous manual alignment at all but which only contains monological narrative texts, we first used WebMAUS to carry out an unconstrained forced alignment and then, if necessary, corrected the resulting boundaries of annotation units by hand and finally reran WebMAUS in constrained alignment mode on the same sessions. In order to ensure the quality of our linguistic analyses, we also decided to manually check the automatic word alignment procuded by WebMAUS using ELAN and to correct larger alignment errors by hand. We did not, however, check each and every word individually but rather listened through the recordings and only corrected the boundaries of clearly misaligned words. In this section, we compare the automatic word-level alignment produced by WebMAUS with the manually corrected version of this alignment.[15] The results reported here thus do not arise from a comparison with an independently created gold-standard alignment, but rather represent a practical evaluation providing information about what percentage of automatically aligned words we deemed to be in need of correction and how far their boundaries needed to be shifted on average.
Table 3 shows the percentage of words in our project corpus and its subcorpora whose boundaries (start time or end time or both) have been manually corrected and by how many milliseconds on average their boundaries were shifted. The latter figure only takes those words into account whose alignment was modified in the manual correc-

$p < 0.001$), Bora easy ($W = 198150.5$, $p < 0.001$), Bora hard ($W = 43653.5$, $p < 0.001$), Even easy ($W = 352639$, $p = 0.009$), Even hard ($W = 135008.5$, $p < 0.001$), German easy ($W = 461580$, $p = 0.029$), and Sri Lanka Malay easy ($W = 83156$, $p = 0.982$).

[11] Baure easy ($W = 667506$, $p < 0.001$), Baure hard ($W = 1145204$, $p < 0.001$), Bora easy ($W = 152971$, $p = 0.013$), Bora hard ($W = 31530.5$, $p < 0.001$), Even easy ($W = 376333$, $p < 0.001$), Even hard ($W = 150200.5$, $p < 0.001$), German easy ($W = 569266$, $p < 0.001$), and Sri Lanka Malay easy ($W = 87779$, $p = 0.1768$).

[12] German vs. Baure easy ($W = 777309.5$, $p < 0.001$), German vs. Bora easy ($W = 285236$, $p = 0.063$), German vs. Even easy ($W = 426543.5$, $p < 0.001$), and German vs. Sri Lanka Malay easy ($W = 223143$, $p < 0.001$).

[13] Constrained alignment: Baure easy vs. hard ($W = 803187.5$, $p < 0.001$), Bora easy vs. hard ($W = 28816.5$, $p < 0.001$), Even easy vs. hard ($W = 158166.5$, $p < 0.001$) all go in the expected direction. Unconstrained alignment: Baure easy vs. hard ($W = 909481$, $p < 0.001$) in the opposite direction, Bora easy vs. hard ($W = 14928$, $p < 0.001$) in the expected direction, Even easy vs. hard ($W = 144785$, $p < 0.001$) in the opposite direction.

[14] Baure easy vs. hard ($W = 622753.5$, $p < 0.001$) in the opposite direction, Bora easy vs. hard ($W = 58058$, $p = 0.129$), Even easy vs. hard ($W = 192561.5$, $p = 0.827$).

[15] At the time of writing of this paper, not all sessions in our corpus were already manually checked so that not all sessions could be used in this evaluation. This problem mostly affects the Texistepec Popoluca and N|uu subcorpora.

Language	Texts	Words per AU	Words	Perc. Corrected Words	Time Shift Mean	Time Shift Median	Time Shift Max
Baure	60	3.72	28,587	19.94%	540 ms	206 ms	15,289 ms
Bora	32	6.94	20,846	68.75%	361 ms	81 ms	15,970 ms
Chintang	52	3.96	46,599	13.67%	2,408 ms	377 ms	18,995 ms
Even	36	6.94	22,917	38.99%	373 ms	147 ms	8,651 ms
N|uu	2	4.54	1,642	20.95%	194 ms	74 ms	3,060 ms
Popoluca	2	3.22	2,269	27.46%	202 ms	65 ms	4,424 ms
Sakha	16	7.30	30,848	38.83%	338 ms	100 ms	10,488 ms
Overall	200	4.95	153,708	31.41%	646 ms	137 ms	18,995 ms

Table 3: Overview of manually corrected word boundaries after constrained forced alignment of the NTVR project corpora

tion stage. Overall, 31.41% of all word alignments were corrected in the manual correction stage (word start or end time or both). This percentage of corrected word alignments varies from only 13.67% for Chintang to 68.75% for Bora. However, it is not clear that it is sensible to compare the figures for two individual languages because the manual correction work had to be carried out by several people in parallel (including the first author, Alena Witzlack-Makarevich, and several student assistants listed in section 5.), who were assigned to different subcorpora, in order to save time. It may simply be the case, for example, that the person correcting one language was a little bit more meticulous than the person correcting a different language. In general, one can probably expect a rate of about 30% manual corrections of word times on a typical language documentation corpus that provides a prealignment usable for constrained forced alignment. An interesting observation is that the percentage of corrected word times exhibits a strong positive correlation with the average size of the prealigned annotation units (AUs) in a language's subcorpus, provided as the mean number of words per annotation unit in the third column of Table 3 ($r = 0.75$, $t = 2.55$, $df = 5$, $p = 0.05$). As one would probably expect, the number of word time corrections that are required increases as the average size of the annotation units (sentence, intonation unit, paragraph, etc.) gets larger. Smaller prealigned annotation units simply contain more information about where words are located and provide less opportunity for the forced alignment to go wrong. A rate of about 30% manually corrected word times also accords with our experience that manually aligning all word boundaries from scratch inside pre-existing larger annotation units takes at least three times more time than manually correcting automatically aligned word boundaries. In our experience, this ratio increases even more as the prealigned annotation units increase in size.

As the last three columns in Table 3 show, the mean size of the time shift of 646 ms between the automatically aligned word start and end times and the manually corrected ones again seems to be somewhat inflated by some rare extreme corrections of more than 10 seconds, which could only occur inside very large annotation units or in case the existing prealignment into annotation units was incorrect in some instances. The overall median time shift of 137 ms is a better characterization of how far word boundaries typically needed to be shifted in the manual correction stage and is also more in line with the median alignment error obtained for constrained forced alignment on the manually aligned test sessions discussed in section 3.1. above.

An interesting observation, albeit a preliminary one because manual word time corrections have only been carried for two of the included texts, is that the results of constrained alignment using WebMAUS obtained for the N|uu subcorpus compare quite favourably with the results for the other languages in table 3, both with regard to a moderate number of corrected word boundaries of 20.95% and with regard to the average time shift in case of word time corrections: The mean time shift of 194 ms and the median of 74 ms for N|uu are way below the overall mean and median time shift of 646 ms and 137 ms, respectively. These results are noteworthy, in our opinion, because the SAM-PA inventory currently provided by WebMAUS lacks a lot of the typical consonants of the N|uu language, particularly different clicks, which therefore had to be mapped to other consonants contained in the WebMAUS SAM-PA inventory based on vague intuitions about acoustic and/or articulatory similarity. A relatively crude grapheme-to-phoneme conversion like the one for N|uu thus does not seem to impede successful forced alignment. Neither the exactness of the grapheme-to-phoneme conversion nor the use of acoustic models trained on the specific language one wants to align therefore seem to be crucial to the success of automatic forced alignment with WebMAUS.

Finally, the two Texistepec Popoluca texts in our corpus in which word times have already been manually corrected allow us to take a look at the average time error between unconstrained alignment and the final word start and end times obtained after manual correction of constrained alignment (this time for all wordtimes not only manually corrected ones); cf. Table 4. At least for these two sessions, unconstrained alignment worked quite well, as the overall median time error of 110 ms for word start and end times shows. But the results of unconstrained alignment vary quite a bit even between these two texts. The median time error of only 30 ms for the text "Pepito" is much smaller than the median time error of 171 ms obtained for "La Chichimeca" (Wilcoxon's test: $W = 609477$, $p < 0.001$). One relevant factor, in addition perhaps to the completeness of the transcription, probably is the overall length of the text that is aligned using unconstrained forced alignment since longer recordings and transcriptions provide more opportunity for confusions over longer stretches of the recording.

Session	Words per AU	Words	Time Shift Mean	Median	Max
La Chichimeca	3.21	2,037	749 ms	171 ms	7,939 ms
Pepito	3.32	232	284 ms	30 ms	6,926 ms
Overall	3.22	2,269	701 ms	110 ms	7,939 ms

Table 4: Comparison between word times obtained using unconstrained forced alignment and manually corrected word boundaries from constrained alignment for two Texistepec Popoluca texts

3.3. Evaluation on an Independently Manually Aligned Corpus of Hoocąk Recordings

The final evaluation of WebMAUS on language documentation data that we carried out within our project is based on a corpus of Hoocąk recordings provided to us by Iren Hartmann. Even though we are hoping to be able to use this corpus also in research on the noun-to-verb ratio, it has been manually time aligned at the word-level independently of our NTVR project. It thus represents a good gold standard test corpus for constrained and unconstrained forced alignment with WebMAUS. As the manual time alignment of words has been carried out in ELAN using a "Time_Subdivision" relation between annotation units and the words contained in them, which does not allow for gaps between words, instead of using the more flexible "Included_In" relation, only word start times are correctly aligned with the audio signal, while word end times automatically coincide with the start of the following word. For this reason, the evaluation results discussed in this section are based on word start times only.

Table 5 provides the results of evaluating WebMAUS on this Hoocąk corpus using both unconstrained and constrained forced alignment. As one would expect, the mean word start and end time difference between constrained forced alignment and manual alignment is much lower than the mean time difference between unconstrained forced alignment and manual alignment: 279 ms (constrained alignment) vs. 4,279 ms (unconstrained alignment). This difference is highly statistically significant according to an unparametric Wilcoxon test ($W = 80786878$, $p < 0.001$). Interestingly, however, the median time differences are much closer together with a medium error of 60 ms in the case of unconstrained alignment compared to a medium error of 50 ms in the case of constrained alignment. This suggests that on the 41 Hoocąk texts we could use for this evaluation, which mostly contain well-transcribed monological narratives, unconstrained forced alignment worked almost as well as constrained forced alignment, except for a few outlier cases of extreme misalignment. The maximal alignment error was over three minutes in the case of unconstrained alignment and over 21 seconds in the case of constrained alignment. The former resulted from an over three minute long but untranscribed stretch of English speech interrupting the transcibed Hoocąk speech. Since no prealigned annotation units were located in this region of the audio file, the resulting misalignments could easily be avoided by the constrained forced alignment. The maximal alignment error in the case of constrained alignment was due to a very long annotation unit containing just two words one of which was misaligned with an untranscribed response from a listener. Such relatively rare cases of extreme alignment errors again inflate the mean time differences so that the automatic alignment results produced by WebMAUS, which we find quite impressive, particularly also those from unconstrained automatic alignment in the case of this Hoocąk test corpus, may look underwhelming at first sight.

4. Discussion and Conclusion

Despite the incomplete inventory of phonemes and acoustic models provided by WebMAUS, our often somewhat crude grapheme-to-phoneme conversion using simple search-and-replace rules, and the use of simple HMM forced alignment only, our experiments with WebMAUS have produced quite promising results. We were able to successfully carry out a word-level alignment of entire transcriptions spanning minutes or even a whole hour of audio with remarkably few serious alignment errors, especially when we could use existing manually aligned annotation units to constrain the automatic alignment, as is usually the case for most recent language documentation corpora. An evaluation on several manually aligned test sessions showed that the median time differences between constrained automatic versus manual alignment are comparable to median human interaligner time differences. Our evaluation studies on corpora of Hoocąk and Texistepec Popoluca recordings also showed that even unconstrained automatic alignment can yield quite impressive alignment results for well-transcribed monological narratives, with median time differences close to those of constrained automatic alignment. This depends, however, on the nature of the transcribed recording and the quality of the transcription. Unconstrained automatic alignment is unable to deal with dialogical texts with many speaker overlaps and can be led astray by parts of the recording that are not transcribed. We have also made the more general observation that the quality of automatic alignment with WebMAUS depends much less on the quality of grapheme-to-phoneme conversion and acoustic models than on the quality and completeness of the transcription. Untranscribed parts of a recording such as filled pauses, backchannel responses, or conversely, transcribed words that do not actually occur in the audio signal will deteriorate the quality of automatic alignment. In the case of constrained forced alignment, we have also observed that smaller prealigned annotation units lead to a better alignment quality.

We believe that the possibility to automatically align transcriptions with audio/video data and to segment the latter into phones and words without the necessity of training acoustic models for individual languages, relying in-

				Unconstrained Alignment			Constrained Alignment		
Language	Texts	Words per AU	Words	mean	median	max	mean	median	max
Hoocąk	41	8.14	12,287	4,279 ms	60 ms	194,289 ms	279 ms	50 ms	21,738 ms

Table 5: Time differences between unconstrained and constrained forced alignment with WebMAUS and a pre-existing manual word-level time alignment of a Hoocąk corpus (word start times only)

stead on accurate transcriptions in combination with larger, manually prealigned annotation units, will enable a wide range of possible research questions for language documentation corpora, including large-scale phonetic studies and corpus-based psycholinguistic studies that investigate the flow of speech through time, which are otherwise only feasible for "major" languages and well-funded long-term corpus building projects. We also believe that WebMAUS is a promising tool for the automatic alignment of heritage corpora when combined with manual correction stages.

5. Acknowledgements

The research for this paper was carried out within the DoBeS project "The relative frequency of nouns, pronouns, and verbs cross-linguistically" funded by the Volkswagen foundation. We would like thank all our colleagues who have contributed their language documentation corpora to our project: Balthasar Bickel, Swintha Danielsen, Tom Güldemann, Iren Hartmann, Sebastian Nordhoff, Brigitte Pakendorf, Robert Schikowski, Frank Seifart, Søren Wichmann, and Alena Witzlack-Makarevich. We are also very grateful to Iren Hartmann for letting us use her manually aligned Hoocąk recordings as a test set. We also would like to give a special thanks to our student assistants who have performed most of the manual alignment and corrections: Helen Geyer, Lena Sell, Lisa Steinbach, Laura Wäggerle, and Evgeniya Zhivotova. Furthermore, we also thank the members of the AUVIS project, Nikolaus Himmelmann, Meytal Sandler, and Volker Unterladstetter for allowing us to use one of their German transcriptions in our evaluation. Last but not least, we are very thankful to the team of the Bavarian Archive for Speech Signals (BAS) in Munich for their great support.

6. References

L. Bombien, S. Cassidy, J. Harrington, T. John, and S. Palethorpe. 2006. Recent developments in the Emu speech database system. In *Proceedings of the Australian Speech Science and Technology Conference in Auckland, New Zealand*, pages 313–318.

S. Burger, K. Weilhammer, F. Schiel, and H. G. Tillmann. 2000. Verbmobil data collection and annotation. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 537–549. Springer, Berlin/Heidelberg.

W. Chafe. 2002. Searching for meaning in language. A memoir. *Historiographia Linguistica*, 29(1/2):245–261.

C. DiCanio, H. Nam, D. H. Whalen, H. T. Bunnell, J. D. Amith, and R. Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America*, 134(3):2235–2246.

N. P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

A. Kipp, M. B. Wesenick, and F. Schiel. 1996. Automatic detection and segmentation of pronunciation variants in German speech corpora. In *Proceedings of the ICSLP in Philadelphia*, pages 106–109.

T. Kisler, F. Schiel, and H. Sloetjes. 2012. Signal processing via web services: The use case WebMAUS. In *Proceedings of Digital Humanities 2012, Hamburg, Germany*, pages 30–34.

F. Schiel, S. Burger, A. Geumann, and K. Weilhammer. 1998. The Partitur format at BAS. In *Proceedings of the 1st International Conference on Language Resources and Evaluation in Granada, Spain*, pages 1295–1301.

F. Schiel, C. Draxler, and J. Harrington. 2011. Phonemic segmentation and labelling using the MAUS technique. Paper presented at the Workshop on New Tools and Methods for Very-Large-Scale Phonetics Research, University of Pennsylvania, January 28-31, 2011.

F. Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the ICPhS in San Francisco*, pages 607–610.

F. Schiel. 2004. MAUS goes iterative. In *Proceedings of the IVth International Conference on Language Resources and Evaluation in Lisbon, Portugal*, pages 1015–1018.

F. Seifart, R. Meyer, T. Zakharko, B. Bickel, S. Danielsen, S. Nordhoff, and A. Witzlack-Makarevich. 2010. Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis. Paper presented at the DobeS Workshop Advances in Documentary Linguistics, Nijmegen, 14-15 October 2010.

F. Seifart. 2011. Cross-linguistic variation in the noun-to-verb ratio: The role of verb morphology and narrative strategies. Poster presented at the Association for Linguistic Typology 9th Biennial Conference, University of Hong Kong, July 21-24, 2011.

J. Strunk. 2013. Automatic alignment of transcriptions and audio for a DoBeS comparative corpus analysis project. Paper presented at the DoBeS Conference Language Documentation: Past – Present – Future, in Hannover, Germany.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation in Genoa, Italy*, pages 1556–1559.