Oral-History.Digital

Automatische Spracherkennung von Oral History-Interviews.

Anleitung für die Nutzung der BAS Web Services

Herdis Kley, Cord Pagenstecher, Florian Schiel unter Mitarbeit von Tobias Kilgus, Peter Kompiel und Philipp Linß

Erarbeitet im Rahmen des Projekts „Oral-History.Digital“, gefördert von der Deutschen Forschungsgemeinschaft (DFG), vom Bayerischen Archiv für Sprachsignale an der Ludwig-Maximilians-Universität München und dem Bereich Digitale Interview-Sammlungen an der Universitätsbibliothek der Freien Universität Berlin

Version 0.6 | 05.09.2022

www.oral-history.digital

Automatische Spracherkennung von Oral History-Interviews.

Anleitung für die Nutzung der BAS Web Services

Ton- und Videoaufnahmen sind die Kernelemente einer Interviewsammlung. Aber um die oft mehrstündigen Interviews durchsuchbar und referenzierbar zu machen, muss die gesprochene Aufzeichnung verschriftlicht werden. Eine manuelle Transkription bringt die besten Ergebnisse, ist aber bei den oft mehrstündigen Interviews sehr zeitaufwändig und kostenintensiv.

Die Weiterentwicklung der softwaregestützten Spracherkennung kann dabei Unterstützung bieten. Sogenannte „Dirty Transcripts“ bieten in den meisten Fällen noch keine lesefähigen Transkriptionen, können aber als Basis für die manuelle Weiterbearbeitung dienen oder bereits für die Volltextsuche genutzt werden.

Um die Audio- und Videodateien durchsuchbar zu machen und die Transkripte (oder Übersetzungen) als Untertitel anzuzeigen, müssen die Transkripte mit Timecodes segmentiert und mit den Mediendateien gekoppelt werden. Auch für dieses sogenannte Alignment gibt es Software-Unterstützung.

Im Projekt *Oral-History.Digital* werden Software-Werkzeuge für die Spracherkennung (Automatic Speech Recognition, ASR) und für die automatische Segmentierung (Alignment) vom Bayerischen Archiv für Sprachsignale an der Ludwig-Maximilians-Universität München geprüft, angepasst und weiterentwickelt. Der Bereich Digitale Interview-Sammlungen an der Universitätsbibliothek der Freien Universität Berlin erarbeitet Handreichungen zur Nutzung dieser Dienste für Oral History-Interviews und prüft die Einrichtung von Schnittstellen aus der Erschließungsplattform *oral-history.digital*. Bitte beachten Sie, dass es aufgrund der Weiterentwicklung beim BAS und den externen Spracherkennern regelmäßig zu Änderungen der Nutzungsbedingungen und Anforderungen kommen kann, so dass diese Anleitung nur den momentanen Stand berücksichtigt und ohne Gewähr ist. Bitte informieren Sie sich vor der Anwendung der BAS Web Services über die aktuellen Nutzungsbedingungen auf deren Webseite: https://clarin.phonetik.uni-muenchen.de/BASWebServices/help/termsOfUsage.

Für die Automatische Spracherkennung (ASR) nutzen Sie direkt die BAS Webservices auf https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface. Die folgende Anleitung unterstützt Sie dabei, für eine Audio- oder Videodatei eines Oral History-Interviews ein timecodiertes Rohtranskript zu erzeugen. Die so erzeugte Untertiteldatei im vtt-Format können Sie in allen gängigen Medienplayern direkt anzeigen lassen, in die Erschließungsplattform *oral-history.digital* importieren oder in einem geeigneten Transkriptionsprogramm, wie zum Beispiel InqScribe, nachbearbeiten.

Die BAS-Webservices bieten für verschiedene Sprachen unterschiedliche Spracherkenner externer Anbieter an. Die größte Abdeckung unterschiedlicher Sprachen bietet die Google-Spracherkennung; andere Anbieter haben andere Vorteile. Auf dem dynamischen Feld der Spracherkennung ändern sich Methoden, Erkennungsraten und Nutzungsbedingungen sehr rasch, so dass alle hier festgehaltenen Empfehlungen stets ohne Gewähr gegeben werden.

Vorbereitung der Mediendateien

Zur Nutzung der BAS-Webservices müssen Sie Mitglied einer akademischen Einrichtung sein oder über CLARIN (https://user.clarin.eu/user/register) einen persönlichen Account beantragen.

Für die automatische Spracherkennung werden die Mediendateien auf **externe Server zur Spracherkennung weitergeleitet**. Vor der Nutzung müssen Sie dafür die (datenschutz)rechtlichen Bedingungen Ihrer Sammlung klären. Bitte beachten Sie sowohl die Nutzungsbedingungen der BAS Web Services als auch der externen Spracherkenner

Größen- und Längenbeschränkungen der verschiedenen Spracherkenner:

Gegebenenfalls müssen lange Mediendateien in kürzere Dateien aufgeteilt werden. Es hat sich auch gezeigt, dass kleinere Dateien (< 1 GB) besser von den Spracherkennern verarbeitet werden.

Die maximale Dateigröße oder Dauer der Medien-Datei unterscheiden sich bei den verschiedenen ASR-Anbietern. Für die **kostenfreie** Nutzung der Spracherkennung (**ohne** Exceed quota key, s.u.) gibt es bei den Anbietern unterschiedliche Begrenzungen:
Beim **Google-Spracherkenner** darf eine Datei maximal 10 min lang sein und insgesamt (alle Nutzer*innen) dürfen pro Monat maximal 16 Stunden verarbeitet werden. Der **Spracherkenner des Fraunhofer-Instituts** hat eine Längenbeschränkung von 5 Stunden (18.000 Sekunden) pro Datei, jedoch dürfen monatlich nicht mehr als 180.000 Sek. (50 h) insgesamt (alle Nutzer*innen) verarbeitet werden und pro Verarbeitung darf eine maximale Dateigröße von 2 GB darf nicht überschritten werden. **Amberscript** bietet eine kostenfreie Spracherkennung von 1 Stunde (3600 Sekunden) im Monat insgesamt (alle Nutzer*innen) an, die einzelnen Dateien dürfen eine Größe von 6 GB und eine Länge von 5 min (300 Sekunden) nicht überschreiten.

Bitte beachten Sie auch die Limits der anderen Anbieter in Bezug auf die Aufnahmedauer und Dateigröße, siehe „Service Manual“ zum ASR-Service: https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/ASR

Exceed quota key (EQK):

Da die kostenfreien Kontingente bei Fraunhofer, Google und Amberscript schnell erschöpft sind, bietet das BAS für diese ASR-Dienste an, eine spezielle BAS-Nutzerlizenz, einen sogenannten „Exceed quota key“ (EQK), zu erwerben (kostenpflichtig!): https://clarin.phonetik.uni-muenchen.de/BASWebServices/help#WhatIfINeedAutomaticTranscriptionOfRecordingsLongerThanAllowedByTheAsrQuotas

Google:
- Kosten: 5 € brutto pro Stunde
- Längenbeschränkung: 3 Stunden (1.800 Sekunden)

Fraunhofer (FhG):
- Kosten: 3,75 € brutto pro Stunde
- Längenbeschränkung: 5 Stunden (18.000 Sekunden)
- Größenbeschränkung: 2 GB (pro Verarbeitung)

Amberscript:
- Kosten: 29 € brutto pro Stunde
- Größenbeschränkung: 6 GB (pro Verarbeitung)

Die Beantragung und die Abrechnung erfolgen über das BAS. Bei Interesse schicken Sie bitte eine E-Mail mit folgenden Angaben an bas@bas.uni-muenchen.de:

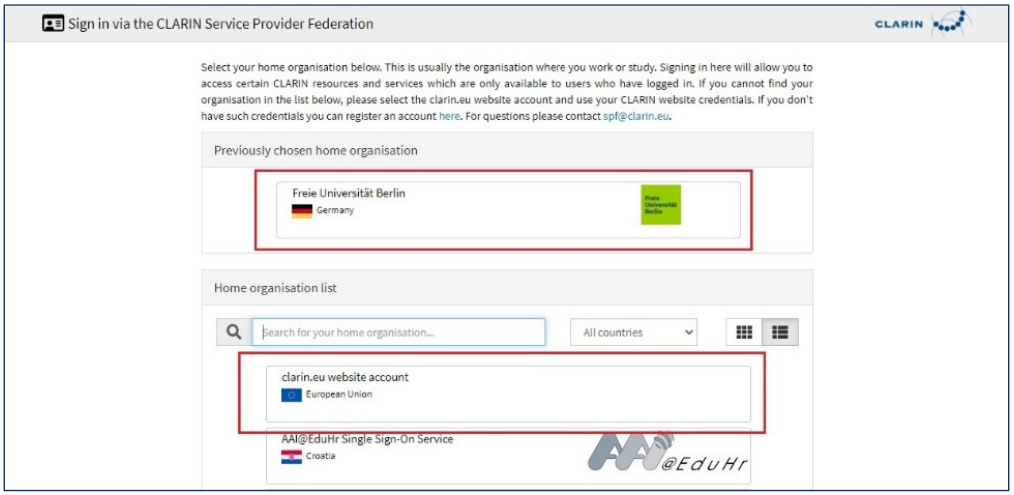
1. Rechnungsadresse (Abrechnung erfolgt alle drei Monate oder auf Wunsch auch vorher)
2. E-Mail-Adresse für Kontakt und Informationen
3. Erster Monat der vorgesehenen Nutzung (z.B. den laufenden Monat)

Um ein schnelles Hochladen und Bearbeiten der Dateien zu ermöglichen, empfiehlt es sich, bei Videodateien die Tonspur herauszuspielen und als WAV-Datei mit mindestens 16kHz und in Mono abzuspeichern. Wenn das nicht möglich ist, kann man auch das Video selber an den BAS-Server schicken (dauert länger und wird dann auf dem Server automatisch transkodiert). Folgende Formate sind erlaubt: aiff, au, avi, flac, flv, mpg, mp3, mpeg, mp4, nis, nist, ogg, snd, sph, wav.

Nachfolgend wird zunächst der komplette Workflow der Spracherkennung erklärt, bei dem die Pipeline inklusive des Subtitle Services durchlaufen wird, und Sie als Ergebnis eine Untertiteldatei erhalten, die Sie direkt in o*ral-history.digital* hochladen können. Falls Sie die Möglichkeit haben möchten, das Transkript im Nachhinein noch anzupassen, wählen Sie in der Pipeline bitte bei Ausgabeformat **BAS Partitur Format (bpf)** und erstellen im Nachhinein über den Subtitle Service die Untertiteldatei (siehe Kapitel „Untertitel/Transkript erstellen“ auf Seite 8). Dies bietet sich an, wenn zum Beispiel nicht nur nach jedem Satz ein Umbruch, also ein neuer Timecode gesetzt werden soll, sondern lange Sätze noch mal geteilt werden müssen.

Spracherkennung (Pipeline with ASR)

1. Starten Sie Google Chrome und gehen Sie auf die Webseite **Pipeline with ASR:** https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/PipelineWithASR



2. Wählen Sie aus der Liste der Institutionen Ihre Heimatorganisation aus. Wenn diese nicht in der Liste enthalten ist oder Sie nicht institutionell angebunden sind, können Sie sich über „clarin.eu website account“ mit Ihrem CLARIN-Account anmelden.

   Bemerkung: Sie werden auf die Anmeldeseite der ausgewählten Institution weitergeleitet.

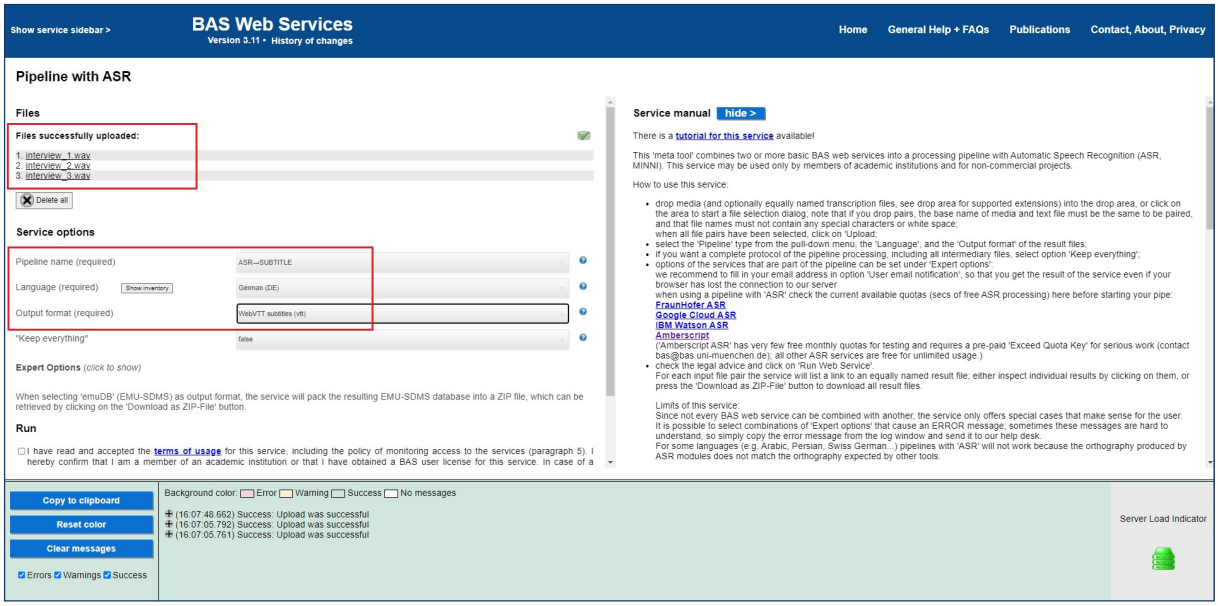
3. Laden Sie die Medien-Datei hoch.

   Bemerkung: Das Hochladen kann je nach Dateigröße einige Minuten bis zu einer Stunde dauern.

   Hinweis: Falls sie gleichzeitig mehrere Dateien verarbeiten möchten, können Sie an dieser Stelle mehrere Dateien hochladen. Bitte öffnen Sie dafür kein weiteres Browserfenster, um parallel mehrere Dateien an die BAS Spracherkennung zu schicken. Bitte beachten Sie, dass Sie für die Nutzung der Sprechererkennung nur eine Einzeldatei über die Pipeline verarbeiten können.

4. Optional: Prüfen Sie für die automatische Sprechererkennung (Diarization) wie viele Sprecher vorkommen, in welcher Reihenfolge sie auftauchen, und welche Sprecherkürzel verwendet werden sollen.

   Beispiel: Zuerst spricht der Interviewer (INT), dann die Zielperson (XX), viel später macht dann der Kameramann eine längere Bemerkung (KAM), dann wäre die Sprecherreihenfolge: INT,XX,KAM



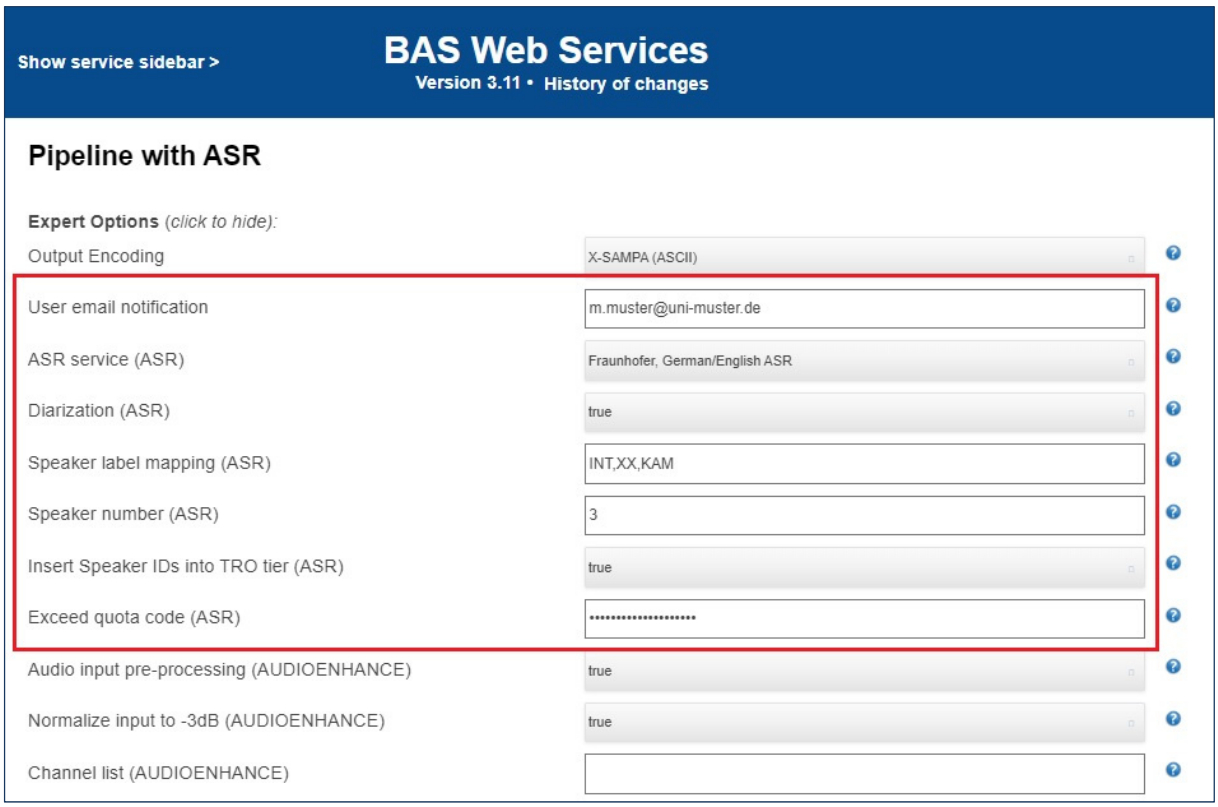
5. Geben Sie unter **Service Options** folgende Werte ein:

   a. Pipeline name: **ASR-Subtitle**

   b. Language: **[Wählen Sie die Sprache des Interviews aus]**

   c. Output format: **WebVTT subtitles (vtt)**

   Bemerkung: Falls Sie die Möglichkeit haben möchten, die Segmentgröße im Nachhinein zu verringern (durch Vorgabe einer maximalen Wortanzahl), wählen Sie als Ausgabeformat **BAS Partitur Format (bpf)** und erstellen Sie die Untertiteldatei in einem zweiten Schritt laut Kapitel „Untertitel/Transkript erstellen“ auf Seite 8.



6. Klicken Sie auf **Expert Options**.

Bemerkung: Es erscheint ein Liste, in der Sie Folgendes ergänzen:

a. Geben Sie bei **User email notification** Ihre E-Mail-Adresse ein.
Bemerkung: Die Arbeit des Web Service kann mehrere Stunden dauern. Sie erhalten eine E-Mail »Your Pipeline Job has finished« mit einem Download-Link zugeschickt (nur für 24h gültig!), wenn der Web Service Ihren Auftrag bearbeitet hat.

b. Wählen Sie bei **ASR service (ASR)** den für Ihre Zwecke am besten geeigneten Spracherkenner aus oder belassen Sie die Vorauswahl bei **Automatic Selection**.
Achtung: Ihre Daten werden im Zuge der Bearbeitung durch die BAS Webservices auf deren Servern (an der Ludwig-Maximilians-Universität München) gespeichert, bearbeitet und automatisch nach 24 Stunden wieder gelöscht. Bei Nutzung eines Spracherkennungs-Dienstes (ASR), werden Ihre Daten zudem an diesen Spracherkennungsservice geschickt, damit dort das automatische Transkript erstellt werden kann.

c. Wenn eine Sprechererkennung durchgeführt werden soll, ändern Sie bei **Diarization** den Wert zu **true**, andernfalls weiter bei g.

Hinweis: Bitte beachten Sie, dass Sie in diesem Fall nur Einzeldateien verarbeiten können und das gleichzeitige Hochladen mehrerer Dateien in einem Verarbeitungsprozess nicht möglich ist.

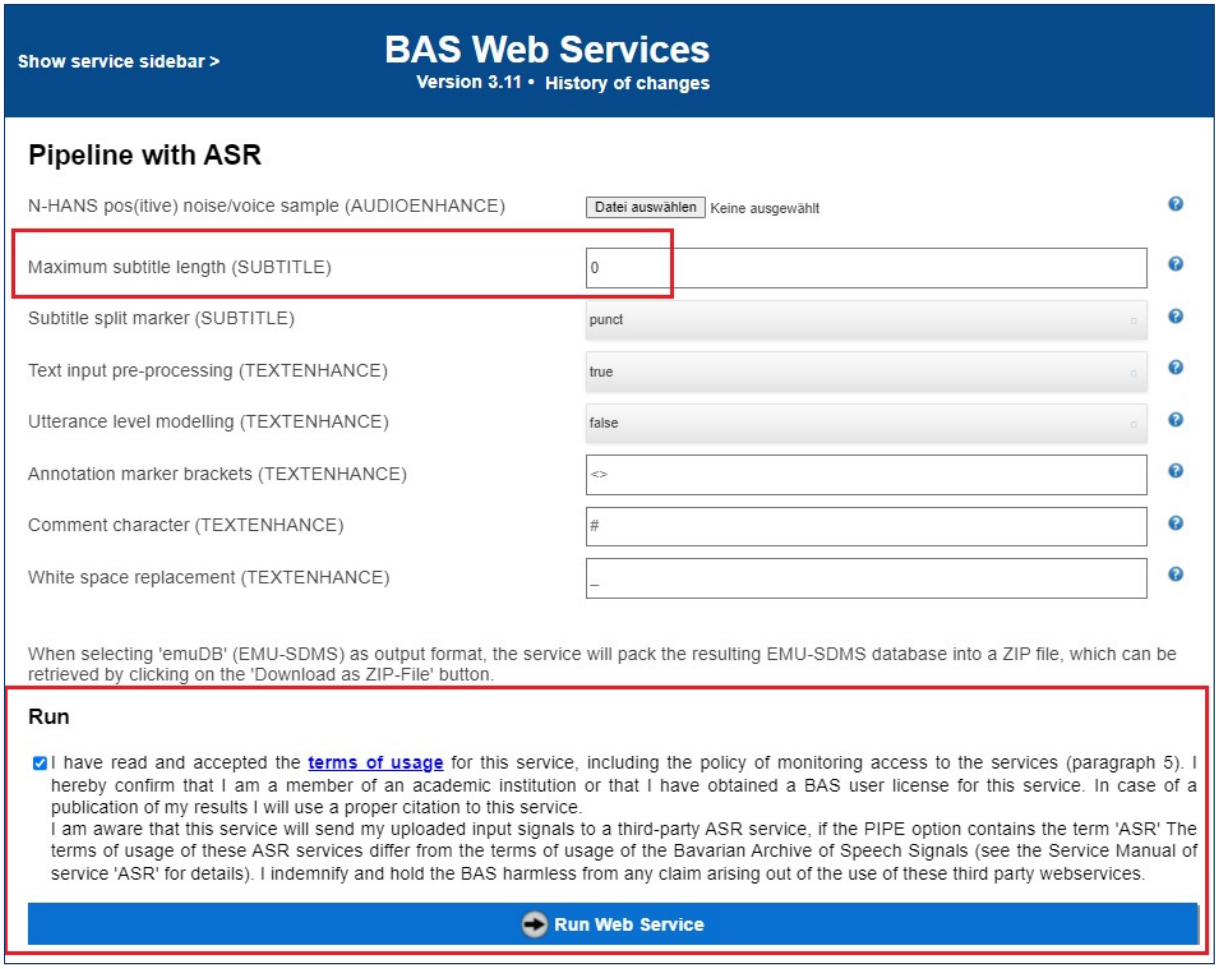
d. Tragen Sie bei **Speaker label mapping (ASR)** die Sprecherreihenfolge ein, z. B. „INT,XX,KAM“. Wenn die Sprecherreihenfolge und -anzahl nicht bekannt sind, weiter bei **f.**
Achtung: Es sind keine Leerzeichen nach dem Komma erlaubt.

e. Geben Sie bei **Speaker number (ASR)** die Anzahl der Sprecher ein, z. B. 3.

f. Ändern Sie bei **Insert Speaker IDs into TRO tier (ASR)** den Wert zu **true.**

g. Geben Sie bei **Exceed Quota Key (ASR)** ggf. den vom BAS zugeschickten Code ein.
Achtung: Bitte stellen Sie sicher, dass Sie bei **ASR service (ASR)** den entsprechenden Spracherkenner ausgewählt haben.

h. Maximum subtitle length: **0**
Bemerkung: Maximum subtitle length: 0 bedeutet, dass neue Segmente erst nach einem abschließenden Satzzeichen (Punkt, Ausrufezeichen, Fragezeichen, Doppelpunkt und Auslassungspunkte), gebildet werden. Wenn Sie zusätzliche Timecodes nach einer maximalen Anzahl von Wörtern benötigen, ändern Sie einfach den Wert (1, 10, 30…).



7. Akzeptieren Sie die **Nutzungsbedingungen** und klicken Sie den Button **Run Web Service** an.

   Bemerkung: Da die Arbeit des Web Service mehrere Stunden dauern kann, können Sie nach Eingabe Ihrer E-Mail-Adresse den Browser bedenkenlos schließen.

8. Downloaden und speichern Sie das Ergebnis durch Rechts-Klick auf den vtt-Ergebnis-Link (.vtt ist die Dateierweiterung für das WebVTT Subtitle Format)

   Bemerkung: Falls Sie bei 5.c als Ausgabeformat das **BAS Partitur Format (bpf)** gewählt haben, erstellen Sie bitte die Untertiteldatei in einem zweiten Schritt, siehe Kapitel „Untertitel/Transkript erstellen“ auf der nächsten Seite.

   *MacOS- und Googlemail-Nutzer*innen beachten bitte den Hinweis am Ende der Anleitung.

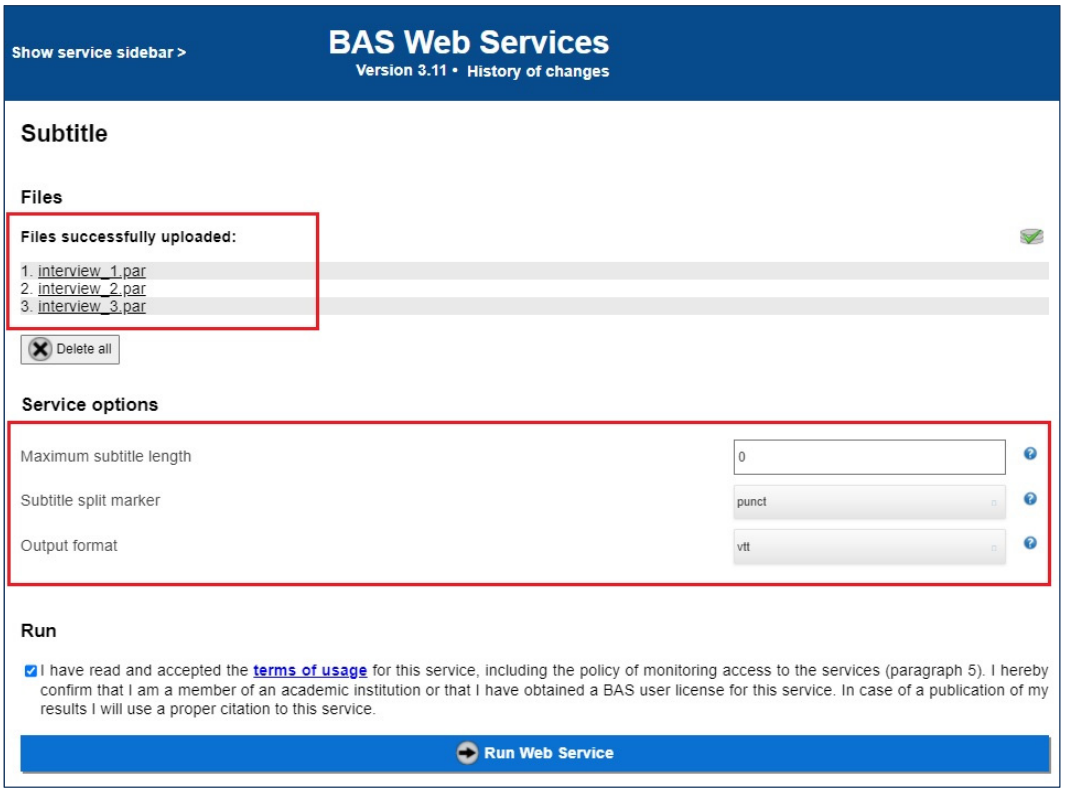
   Achtung: Wenn Sie eine E-Mail ohne Ergebnislink, aber mit einer „Error“-Meldung erhalten, leiten Sie diese E-Mail bitte an die Entwickler des BAS weiter (webservices@bas.uni-muenchen.de).

Die so erzeugte Untertiteldatei im vtt-Format können Sie in allen gängigen Medienplayern direkt anzeigen lassen oder in die Erschließungsplattform von *oral-history.digital* importieren. Für die Nachbearbeitung in einem Transkriptionsprogramm nutzen Sie bitte die **Anleitung „Import und Bearbeitung von vtt-Dateien in InqScribe**“.

Untertitel/Transkript erstellen

Achtung: Dieser Verarbeitungsschritt ist nur notwendig, wenn Sie in der Pipeline das **BAS Partitur Format (bpf)** gewählt haben.

1. Starten Sie Google Chrome und gehen Sie auf die Webseite **Subtitle**: https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Subtitle
2. Laden Sie die PAR-Datei(en) hoch.
3. Geben Sie unter **Service Options** folgende Werte ein:
   a. Maximum subtitle length: **0**
      Bemerkung: Maximum subtitle length: 0 bedeutet, dass neue Segmente erst nach einem abschließenden Satzzeichen (Punkt, Ausrufezeichen, Fragezeichen, Doppelpunkt und Auslassungspunkte) gebildet werden. Wenn Sie zusätzliche Timecodes nach einer maximalen Anzahl von Wörtern benötigen, ändern Sie einfach den Wert (1, 10, 30…).
   b. Output format: **vtt**
      Bemerkung: Sie können zwischen folgenden Ausgabeformaten wählen: srt, sub, vtt, bpf+trn



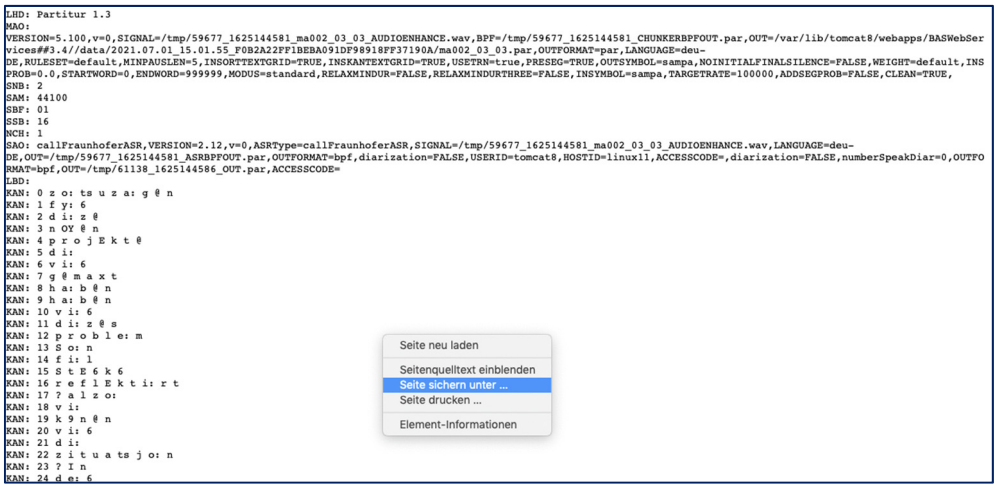
4. Akzeptieren Sie die **Nutzungsbedingungen** und klicken Sie den Button **Run Web Service** an.
   Bemerkung: Der Subtitle Service benötigt i.d.R. nur einige Sekunden/Minuten.
5. Downloaden und speichern Sie das Ergebnis durch Rechts-Klick auf den VTT-Ergebnis-Link. *MacOS- und Googlemail-Nutzer*innen beachten bitte den Hinweis am Ende der Anleitung.

Die so erzeugte Untertiteldatei im vtt-Format können Sie in allen gängigen Medienplayern direkt anzeigen lassen oder in die Erschließungsplattform von *oral-history.digital* importieren. Für die Nachbearbeitung in einem Transkriptionsprogramm nutzen Sie bitte die **Anleitung „Import und Bearbeitung von vtt-Dateien in InqScribe**“.

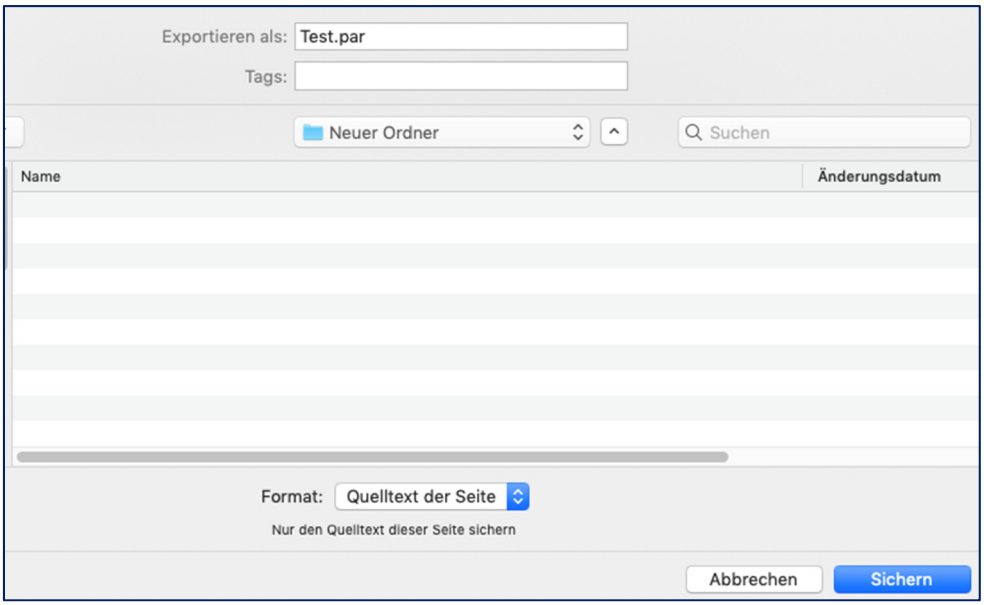
***Hinweis für MacOS- und Googlemail-Nutzer*innen:**

In Apple Mail lässt sich die Datei nicht über einen Rechts-Klick herunterladen. Daher muss man den Quelltext der verlinkten Webseite über den Safari-Browser herunterzuladen.

Dafür wählt man auf der Webseite Rechtsklick -> **Seite sichern unter**



Als Format **"Quelltext der Seite"** wählen und als Dateiendung **.par** (bzw. **.vtt)** eintippen.



Wichtig: bei der folgenden Meldung **"nicht anfügen"** auswählen.