Speech-to-text: Linking Social Survey and Linguistic Infrastructures through speech interviews

SSHOC Workshop

16 April 2021, Online

Speakers

Judith Koops, Generation and Gender Program, NIDI, The Hague

Joris Mulder, LISS Panel

Henk van den Heuvel, Oral History Team – Radboud University, Nijmegen

Giovanni Borghesan, EVS – Tilburg University

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Project:

SSHOC
social sciences & humanities open cloud

Horizon 2020
European Union Funding
for Research & Innovation

Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 45
(20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Duration: 40 months
(January 2019 – 30 April 2022)

Project budget:
€ 14,455,594.08

Project website:
www.SSHopencloud.eu

Objectives:

- creating the social sciences and humanities **(SSH)** part of European Open Science Cloud **(EOSC)**
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

Speakers

Judith Koops
Postdoctoral Researcher
Generation and Gender Program, NIDI, The Hague

Joris Mulder
LISS Coordinator and researcher
LISS Panel

Henk van den Heuvel
Senior Researcher
Oral History Team - Radboud University, Nijmegen

Giovanni Borghesan
Junior Researcher
EVS – Tilburg University

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Agenda

- Introduction
- Overview and description of the project
- Applying Speech to Text software in the Dutch LISS panel
- Processing and analysis of Audio data

- Breakout-rooms
  - Survey developer perspective – Joris Mulder
  - User perspective – Judith Koops
  - Information extraction perspective – Henk van den Heuvel
  - Methodological / analytical insights – Giovanni Borghesan

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Housekeeping Rules

- Keep your microphone muted during other's speeches
- If you have questions, please submit them in the chat
- Presentations will be recorded. Video and slides will be available after the workshop.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Linking Social Survey and Linguistic Infrastructures

Voice recorded interviews

Dr. Judith Koops

Generations and Gender Survey
Netherlands Interdisciplinary Demographic Institute

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

The start of the project

European Open Science Cloud (EOSC)

Social Sciences & Humanities Open Cloud (SSHOC)

Social Sciences Cluster

Humanities Cluster

GGP 2020

European Values Study

CLARIN

cessda

SHARE
SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE

European Social Survey

E-RIHS
EUROPEAN RESEARCH INFRASTRUCTURE FOR HERITAGE SCIENCE

DARIAH-EU

SSHOC
social sciences & humanities open cloud

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Social Sciences & Humanities

Social Sciences Cluster

European Values Study

GGP 2020

- Quantitative
- Structured
- Cross-national representative

Humanities Cluster

CLARIN

- Qualitative
- Complex and rich
- One language / source

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Bringing Social Sciences Cluster & Humanities Cluster together

Collect audio recordings in social survey

Social Science infrastructure Collects data
Humanities infrastructure Processes data

Hypothetical situation

- Audio data collected with GGP (24 countries * 7000 individuals)
- Topics: Relationship types
- Data transcription and sentiment analysis by CLARIN

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Bringing Social Sciences Cluster & Humanities Cluster together

Linguist

- 168,000 language inputs
- Examine same language across countries
- Examine minority languages
- Spread of dialects
- Wealth of socio-economic and demographic

Sociologist / Demographer

- Open-ended questions
- Data quality
- Sentiment towards topic
- Language fluency

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

First things first

Experiment – LISS Panel

- Module 5 topics:
  - Democracy
  - European Union
  - Trust
  - Marriage
  - Unmarried cohabitation
- Combine CAWI and CARI
- Gross N = 400 / Net N = 50 (+)
- Speech to text transcript
- Natural Language Processing tools

Start data collection: this week
Data release: end of 2021 via LISS repository

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

“Can you hear what I’m saying?”

Applying Speech to Text software in the Dutch LISS panel

Challenges and solutions of audio responses in a web survey

Joris Mulder
LISS Coordinator, CentERdata
Tilburg University

SSHOC Speech-to-text Workshop
April 16, 2021

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

What is the LISS Panel?

- Online panel of 5,000 households, comprising of 7,500 individuals (16+)
- Online interviews as method:
  - Probability sample drawn by Statistics Netherlands, from address sampling frame
  - Includes households without Internet access: equipment provided (SimPC)
- Questionnaires each month, 30 minutes
- Incentive € 15 per hour (= € 2,50 for 10 min. survey)

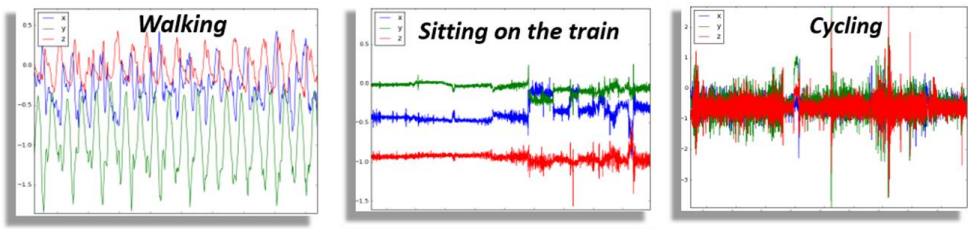
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Advanced measuring devices

- Internet bathroom scales
  - *weight & fat percentage*

- Smartphones
  - *time use, travel behavior / mobility*

- *Accelerometers*
  - *physical activity and pattern recognition (machine learning)*



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Current innovation: Speech To Text

- Due to flexible infrastructure of LISS panel one can easily accommodate new

  inventions:

NIDI Speech to Text study

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Issues to consider

- What software to use?
  - Existing software, open source, develop in-house
- Technical:
  - Integrating in panel software
  - Suitable devices respondents (microphone, headset, etc.)
- Privacy & security:
  - Storing audio files, cleaning the data, is a voice personal data?
- Methodological:
  - Selective non-response (cannot / won't participate)

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Software and technical integration

- External Speech to Text software
- SaaS tool Questfox
- Integrated in the LISS panel

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Privacy & security

- Storing audio files
  - Dedicated SFTP server
- Cleaning data
  - S2T Transcripts
  - Audio files time consuming process
- Privacy
  - Informed consent
  - Instruction: do not mention personal info
  - Respondents voice: personal data?
  - Use of one time respondent ID

SFTP
LISS

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Methodological implications

- Selective non-response
  - Refusals
  - Cannot participate (e.g. no microphone available)
  - Technical issues (e.g. microphone not functioning)
  - Not supported technology

- Discussion
  - How to solve these implications?

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Live demo

- Live demo of the current study

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Linguistic analysis of audio recorded social surveys

By means of language and speech technology

Dr Henk van den Heuvel; CLST,
Radboud University,

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Overview

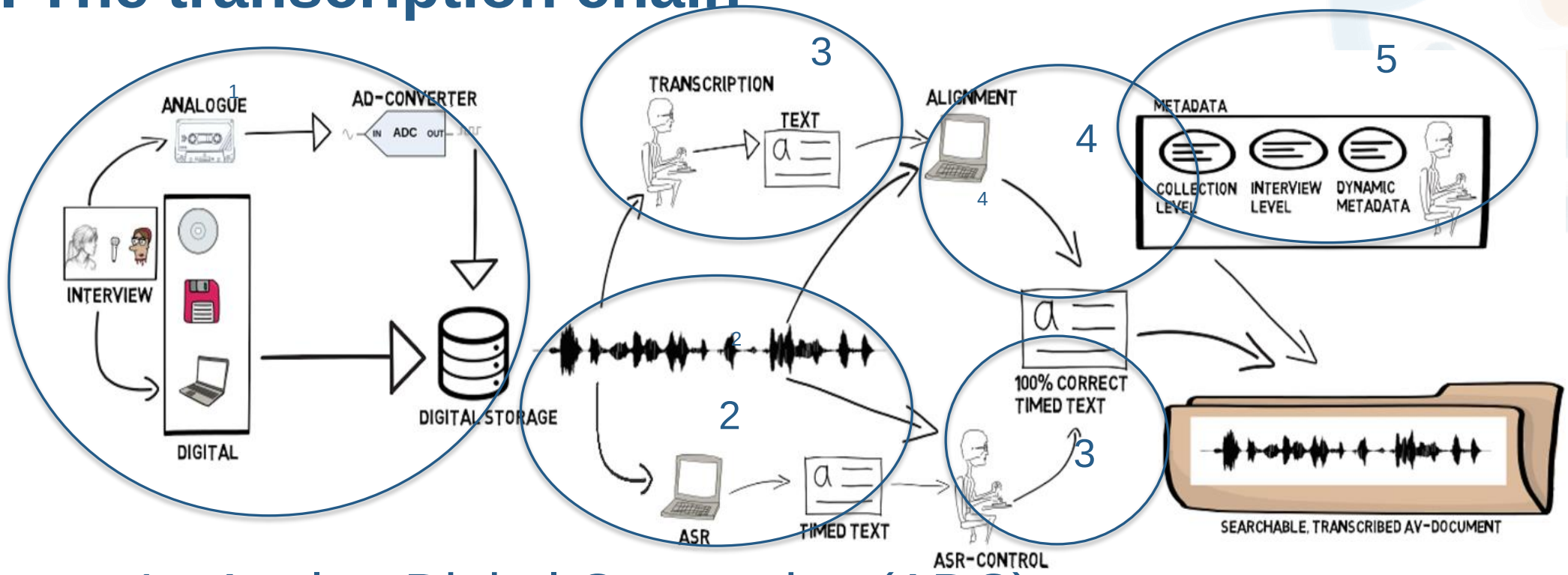
1. Towards transcriptions of audio recordings via automatic speech recognition (The T-Chain)
2. Additional acoustical analysis of the speech signal
3. Text analysis methods

In CLARIN infrastructure: https://www.clarin.eu/

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

1. The transcription chain



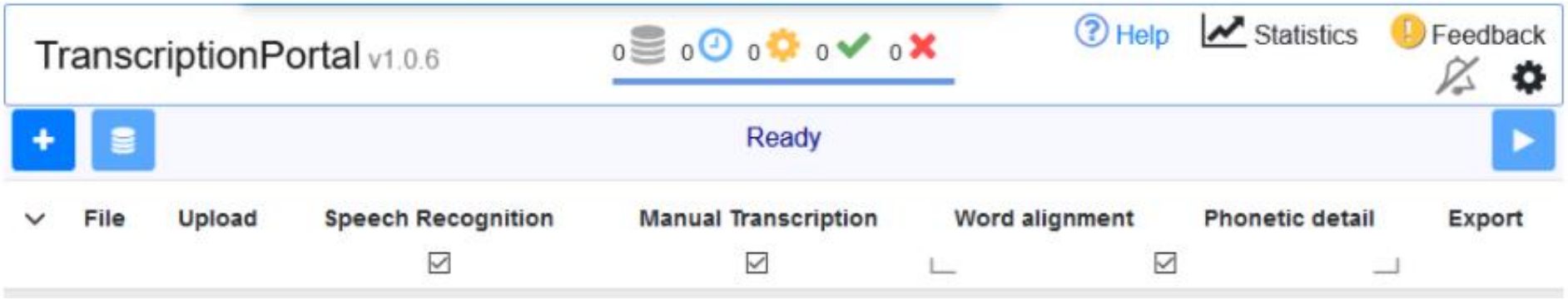
1. Analog-Digital Conversion (ADC)
2. Automatic Speech Recognition (ASR)
3. Transcription (correction)
4. Synchronisation of Audio & transcription (Alignment)
5. Metadata

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

1. The T-Chain

Implemented for CLARIN: https://clarin.phonetik.uni-muenchen.de/apps/TranscriptionPortal/



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

1. T-chain: Automatic speech recognition for survey recordings

1. Avoid speaker overlaps √
2. Use close talk mics ≈√
3. Avoid background noise ≈√
4. Articulate normally

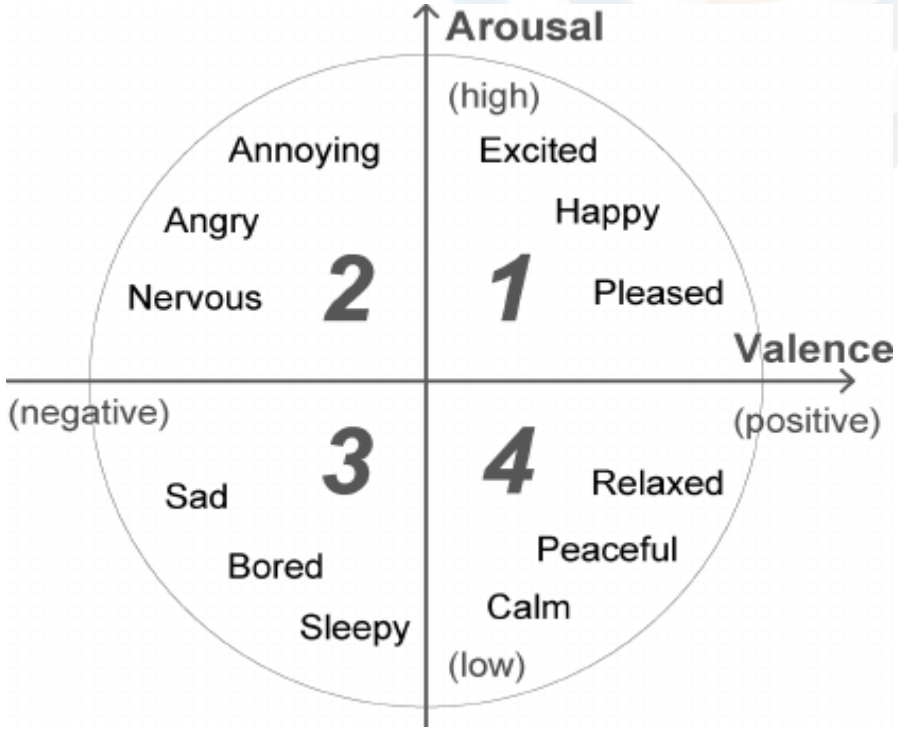
1. Good internet connection

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

2. Additional acoustical analysis

1. Emotional analysis
2. Pauses and volume as indicators of negative valence
3. Pitch variation as indicators of positive valence and higher arousal



Nazareth, D. S., Tournier, E., Leimkötter, S., Janse, E., Heylen, D. K. J., Westerhof, G. J., & Truong, K. P. (2019). An acoustic and lexical analysis of emotional valence in spontaneous speech: Autobiographical memory recall in older adults. *Proceedings of Interspeech*, 3287–3291.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

3. Text analysis

1. Qualitative data analysis
2. Corpus analysis / computational linguistics

See https://speechandtech.eu/tech-and-tools/

e.g. topic modeling: tools for finding a group of words (i.e topic) from a collection of documents that best represents the information in the collection.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

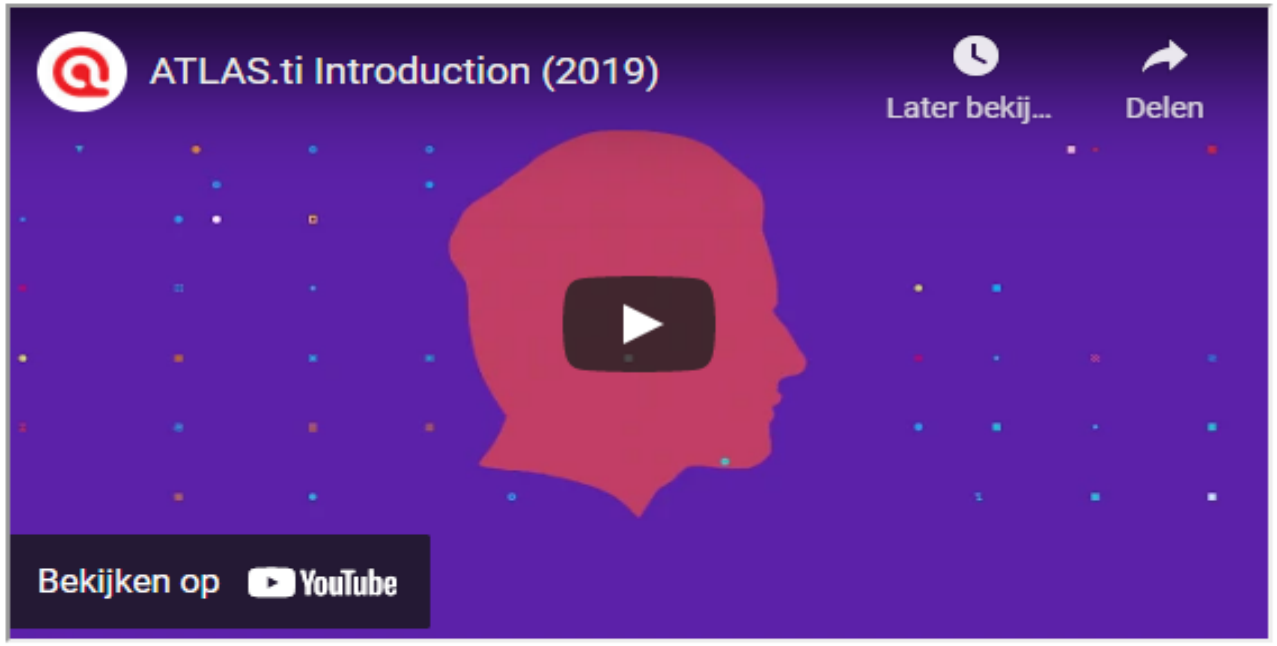
3. Text analysis: Qualitative data analysis

ATLAS.ti | MaxQDA | NVivo | Praat

QDA Workbench - Mixed method research - Many media types

Paid - Installer (Windows, Mac) or web-based (beta)

https://atlasti.com/



ATLAS is a workbench suited for the qualitative analysis of large bodies of textual, graphical, aduio and video data. This paid tool is a standard in many academic institutions, as it allows for many data types, and has a modern and accesible user interface, reminiscent of Microsoft Office.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

3. Text analysis: Qualitative data analysis

ATLAS.ti | MaxQDA | NVivo | Praat

Speech analysis - Phonetics - Open Source

Free - Installer (Windows, Mac, Linux, Chromebook)

https://praat.org/

Praat is not a classic CAQDAS, but an open-source phonetics program containing tools for the recording, analysis, measurement and editing of speech sounds. While the software can be intimidating for new users, Praat is a powerful tool for phonetics research.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

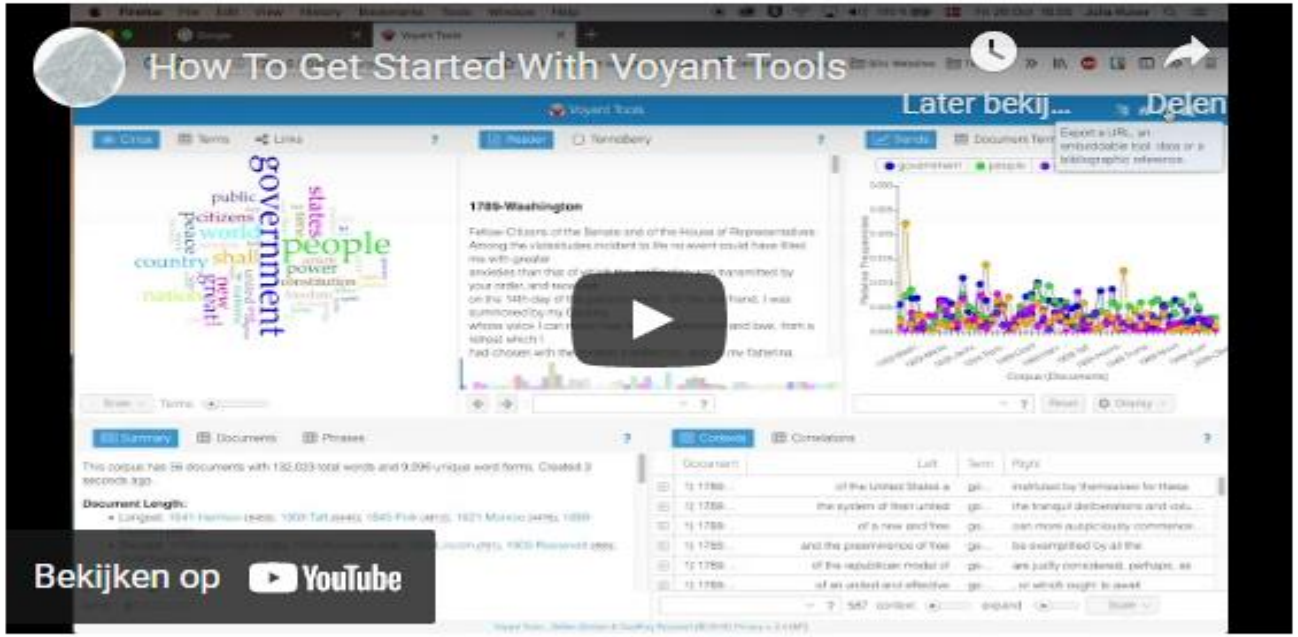
3. Text analysis: Corpus analysis

Voyant Tools | Sketch Engine | Autosummarizer

Online Text-mining - Text analysis - Lightweight, simple

Free - Webbased

https://voyant-tools.org/



Voyant is a lightweight web-based text mining tool which is split up in multiple windows, all having it's own uses when it comes to bulk text analysis. While the tool will not be very useful to generate complete research results, it can be useful for finding trends and commonly named subjects in digital transcripts.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

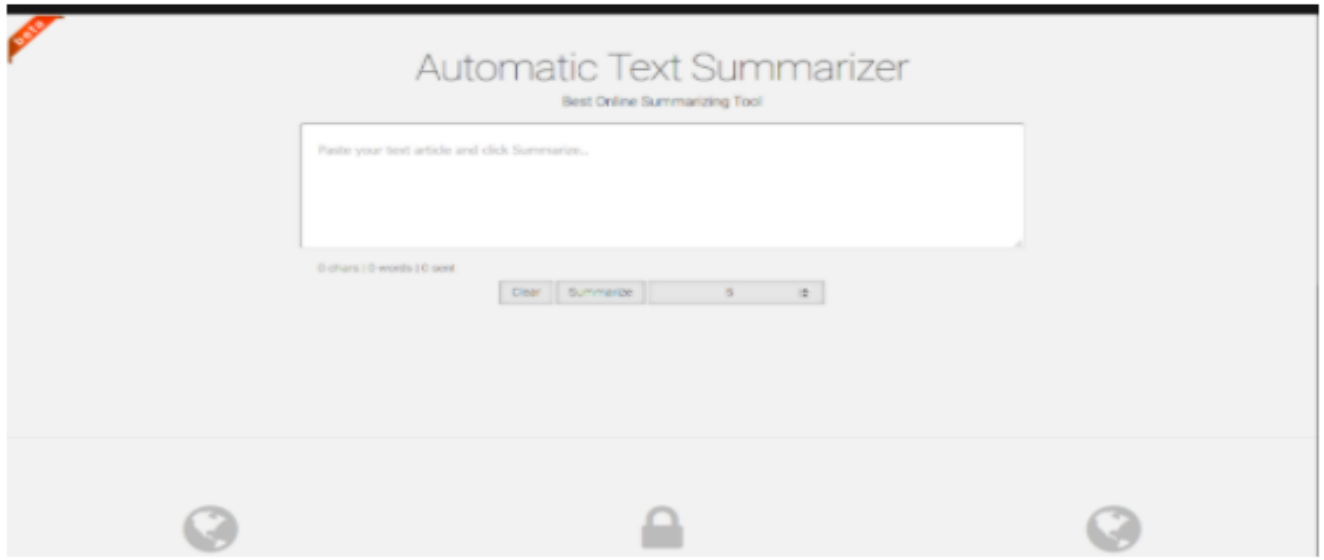
3. Text analysis: Corpus analysis

Voyant Tools | Sketch Engine | Autosummarizer

Automatic summarizer - Artificial Intelligence - Easy-to-use

Free - Webbased

https://www.autosummarizer.com/



Advancements in machine learning and artificial intelligence have caused great leaps in the abilities of algorhythms to automatically summarize texts into a shorter version, selecting which parts are core information and which parts to leave out. Many of such algorhythms exist, and Autosummarizer is just a simple web-based example. While most useful for newspaper articles, the technology in general could be useful for shortening transcripts and finding the common thread of an interview.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Break Out Rooms
20 minutes

- ROOM 1: Survey developer perspective – Joris Mulder
- ROOM 2: User perspective – Judith Koops
- ROOM 3: Information extraction perspective – Henk van den Heuvel
- ROOM 4: Methodological / analytical insights – Giovanni Borghesan

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Plenary Discussion

Speech-to-text: Linking Social Survey and Linguistic Infrastructures through speech interviews

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Thank you for your attention!

Join our community

https://www.sshopencloud.eu

@SSHOpenCloud

info@sshopencloud.eu

/in/sshopencloud

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud