

Archiving and Preservation of Oral History: Recorded Voices of the Past in the Digital Age

A Panel at the DARIAH Annual Event 2025

Marijn Braam, Norah Karrouche, Annette Langedijk, Sanneke Stigter, Jetze Touber



Meaningful Memories: an automated oral history | annotation pipeline

Annette Langedijk (SURF)

Amsterdam Diaries Time Machine



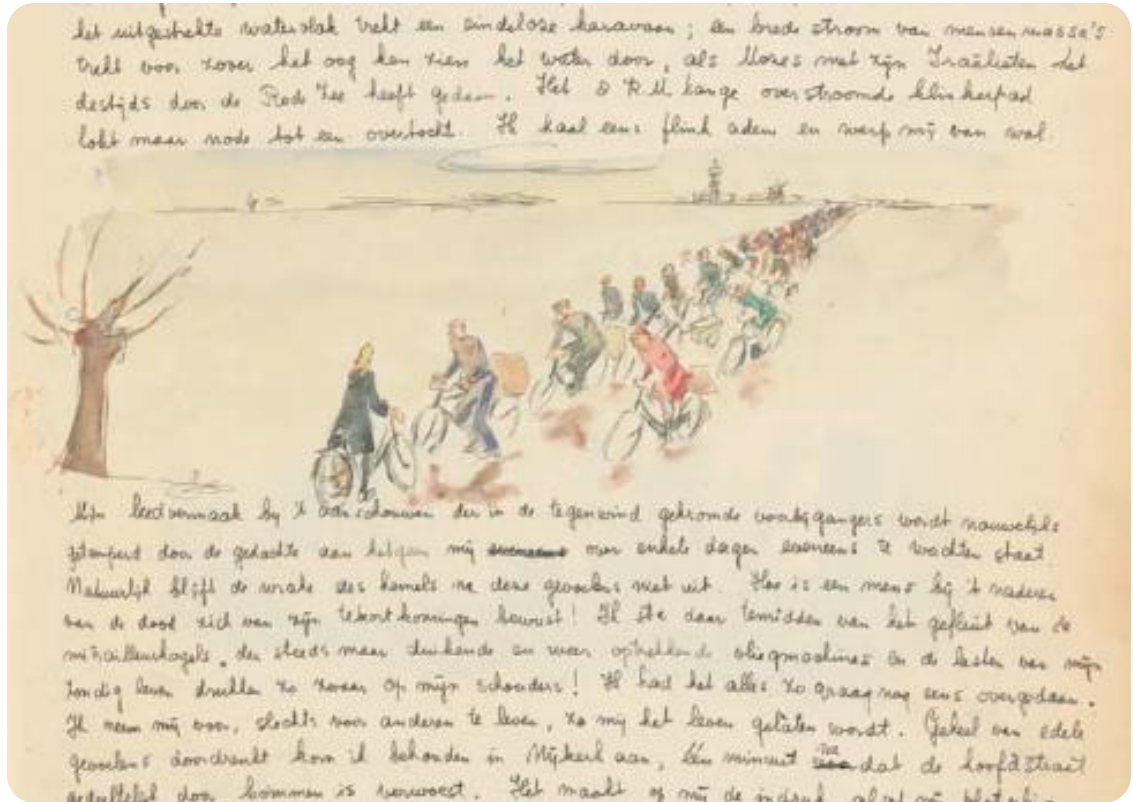
UNIVERSITEIT VAN AMSTERDAM

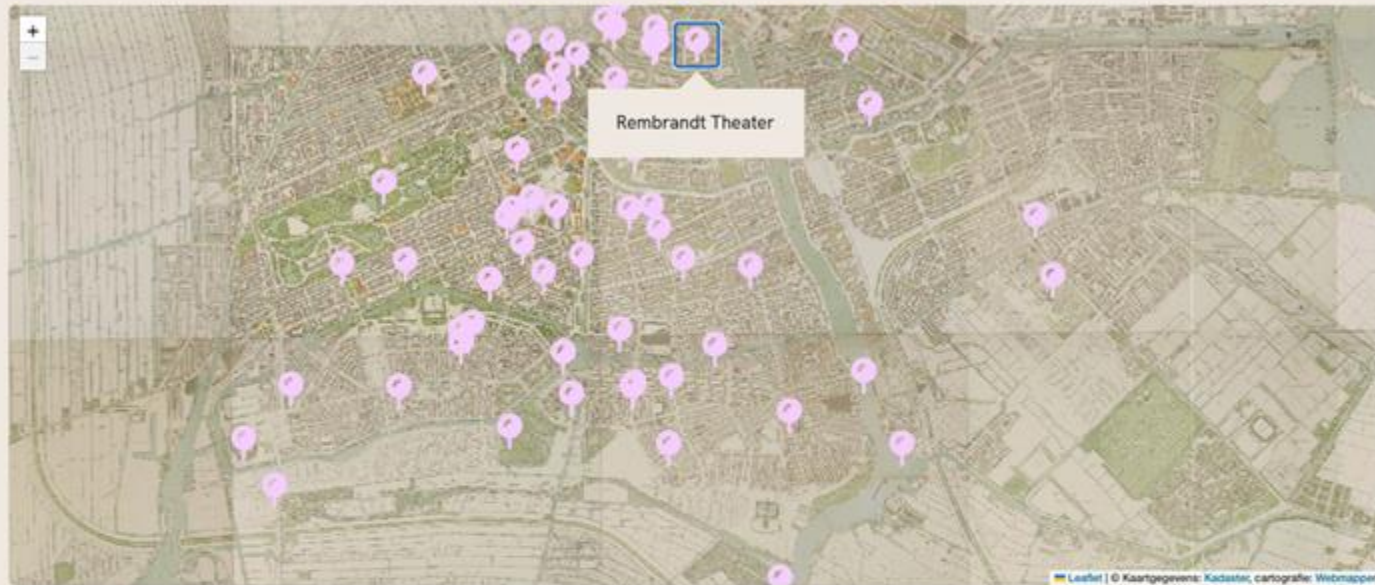


VRJE
UNIVERSITEIT
AMSTERDAM

<https://diaries.amsterdamtimemachine.nl/>

World War II through the eyes of women





Dagboekteksten over Rembrandt Theater

Dagboek Berdi, 1942-1945 1942/1945

desondanks zijn we vanmiddag
of liever gezegd tegen de avond
gaan wandelen. Greet had nog
nooit gezien hoe of het

Dagboek Berdi, 1942-1945 1942/1945

mijn buikje liggen en dan til ik
ieder ogenblik mijn hoofdje op.
[Het Rembrandt theater](#) hier in
Amsterdam is vannacht

Dagboek Berdi, 1942-1945 1942/1945

laden. En nu is de tram nog wel
omgelegd met het oog op de
brand van het [Rembrandt
theater](#). De mensen hangen

Alle organisaties

Ontdek de diverse perspectieven over specifieke organisaties die terugkomen in diverse dagboeken.

42 organisaties



AFC Ajax

Nederlandse profvoetbalclub uit Amsterdam

[Bekijk meer over AFC Ajax](#)



Algemeen Nederlands
Persbureau

het grootste Nederlandse persbureau

[Bekijk meer over Algemeen Nederlands...](#)



Anti-Kominternpact

pact between Nazi Germany and Japan prior to World War II

[Bekijk meer over Anti-Kominternpact](#)



April-meistakingen

April May strikes

[Bekijk meer over April-meistakingen](#)



Barlaeus Gymnasium

[Bekijk meer over Barlaeus Gymnasium](#)



Burgerziekenhuis

voormalig ziekenhuis in Amsterdam

[Bekijk meer over Burgerziekenhuis](#)



Centraal Theater

[Bekijk meer over Centraal Theater](#)



City Theater

[Bekijk meer over City Theater](#)



De Bijenkorf

warenhuis in Amsterdam

[Bekijk meer over De Bijenkorf](#)



De Bijenkorf



De Telegraaf

Nederlandse krant



De Volewijckers

voetbalclub uit Nederland

× Gemeente
× Amsterdam
× Stadsarchief

Joods
Museum

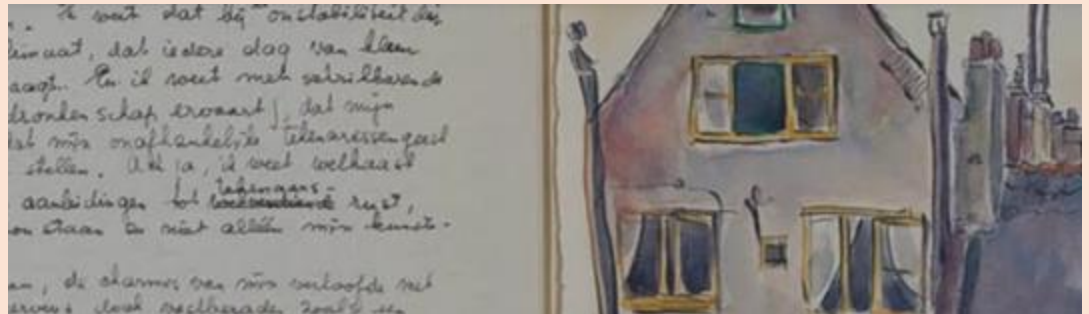
VERZETS
RESISTANCE
MUSEUM

atria

Goal 120 hr innovation project: a Proof-of-Concept

- Extending the Amsterdam Diaries Time Machine with oral history
- Generate fragment-level annotations
- Reducing the amount of manual work

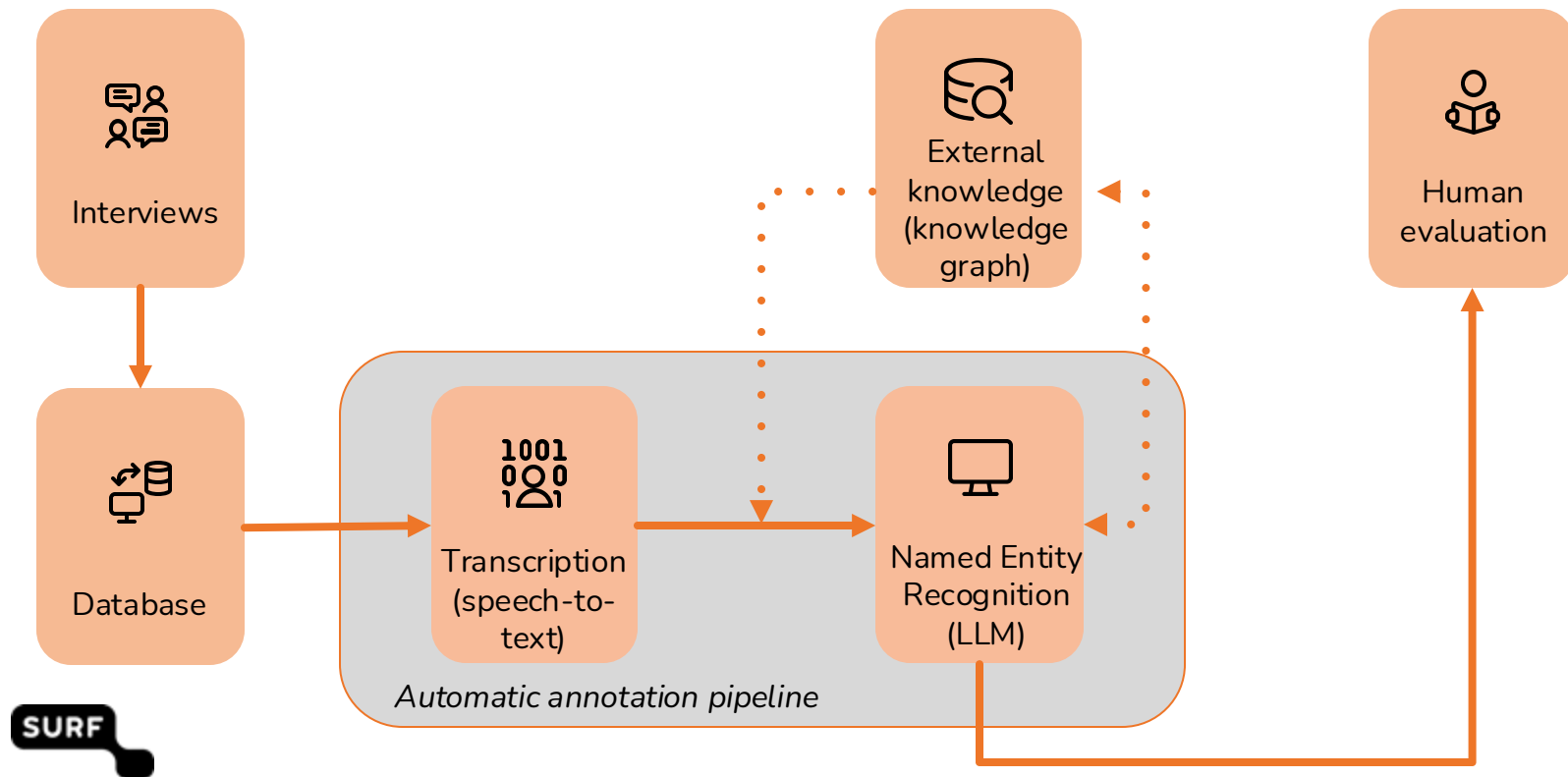
Maintain data provenance!



Modular pipeline

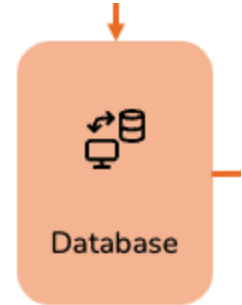


LLM-based transcription and annotation pipeline



Data

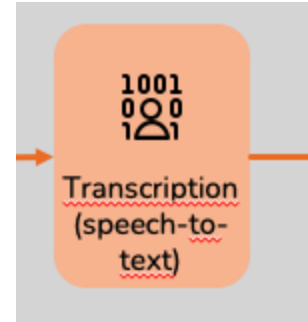
Amsterdam documentaries
by students University of Amsterdam
(master Public History)



Videos (177 files, all in Dutch)

















| Step 1: Transcription Speech-to-text



Whisper and WhisperX
on premise (laptop or SURF AI-hub)



Difference Between **Whisper** and **WhisperX**

Feature	Whisper (by OpenAI)	WhisperX (community extension) 
 Core	Based on OpenAI's Whisper model	Built on Whisper, with added tools
 Word-level timestamps	 Only per sentence or segment	 Yes — accurate word-level timestamps
 Speaker Diarization	 Not included	 Supports speaker identification
 Speed	Decent, but slow for long audio files	Optimized with batching and GPU support
 Timestamp alignment	 Not optimized	 Improved using Wav2Vec2 forced alignment
 Output formats	JSON, plain text	 Extended: JSON + word-level + speaker tags

| What is Named Entity Recognition (NER)?

Text pre-processing (“tokenisation”)

['Barack', 'Obama', 'was', 'born', 'in', 'Hawaii', 'in', '1961', 'and', 'became', 'President', 'in', '2009', '.']

Feature Extraction

Capitalization; Surrounding words (context); Word embeddings; Position in sentence

Entity Detection and Classification

Each token (or group of tokens) is labelled with a NER tag (e.g., PER = person, LOC = location, ORG = organization)

```
json
[
  { "text": "Barack Obama", "type": "PERSON" },
  { "text": "Hawaii", "type": "LOCATION" },
  { "text": "1961", "type": "DATE" },
  { "text": "2009", "type": "DATE" }
]
```

| Step 2: Named Entity Recognition with GLiNER

Advantages:

1. Custom Entity Types (Zero-shot NER)

Ask GLiNER to find things like "emotion", "illness", or "travel plan"

2. Multilingual Support

3. No Training Required

It works out-of-the-box — no model fine-tuning

👑 GLiNER: Generalist and Lightweight Model for
Named Entity Recognition

| Step 3: Theme/Concept extraction with LLM

Focus on:

- overarching themes
- abstract concepts

How:

- Prompt model to extract themes on paragraph-level
- Aggregate results to a top-5

LLM: Llama3



SURF

On premise (SURF AI-hub or laptop)

Artificial intelligence

Direct to →

[AI in research](#)[AI in education](#)[AI in operations](#)[Responsible AI](#)**[GPT-NL](#)**[Dutch AI facility](#)

GPT-NL

SURF is working with TNO and NFI to develop its own open language model called GPT-NL. For strengthening and preserving our digital sovereignty.

About

Non-profit organisations TNO, NFI and SURF are joining forces to develop this model. In doing so, they are taking an important step towards transparent, fair and verifiable use of artificial intelligence, with respect for Dutch and

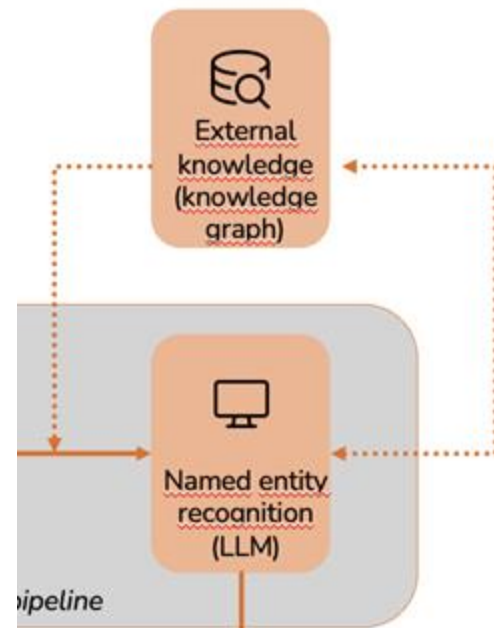
GPT-NL

Link to external Resources

[Wikidata.org](https://www.wikidata.org)

And Dutch-specific:

- [Network of Terms](#) (from Network Cultural Heritage)
- [Cultural Heritage Thesaurus](#) (from Dutch Cultural Heritage Agency)
- [Adamlink.nl](https://adamlink.nl) (reference data for Amsterdam collections)



Human evaluation



Bouw mee aan de digitale tijdmachine van Amsterdam

Datasprint Amsterdam Time Machine in het kader van de ToekomstTiendaagse

De Universiteit van Amsterdam brengt de geschiedenis van de stad tot leven en jij kunt helpen! Tijdens een

Label Studio



Results: WhisperX and GLiNER Named Entity Recognition

‘Daarom is dit ook mijn favoriete plek.’ Wanneer hij mij vertelt over de rellen in de **Spuistraat** van toen, moeten we ons gesprek onderbreken.

Het is nu een onschuldige vuilniswagen die het kabaal veroorzaakt.

‘We zijn echt op locatie he!’ Dit is exemplarisch voor de grote facelift die **Amsterdam** heeft ondergaan.

Kraken is verboden, het is veiliger en schoner op straat en de gebouwen zijn goed onderhouden.

‘**Amsterdam** is nu een **slagroomtaart** die helemaal puntgaaf is.’ Maar diep verborgen in die taart zitten nog steeds dezelfde problemen: woningnood, hoge huurprijzen en speculatie.

In de **jaren 70** is de buurt een no-go area.

Met zijn eigen wetten, zijn eigen regels.

Voor de politie is het dweilen met de kraan op.

Serveren op de **Zeedijk**, dan ging je die kant uit.

En dan kwam je bij de kop van de **Zeedijk** terecht.

En dan liep je daar en dan was het dat je tegen toeristen zei... pas wel even goed op je tas, want zakkenrollers hadden daar hun werkterrein.

De dealers die dan allemaal van die kleine propjes hadden met **heroïne** erin.

Als we dan zagen dat is de pakjesdrager, dan probeerden we die te benaderen.

En dan greep hij hem.

En dan was het zakken leeg of de plek al of mee naar het bureau.

Results: LLM based Theme/Concept extraction

> De Kinkebuurt werd gebouwd in de late negentiende eeuw, samen met een hele reeks andere buurten, zoals de Dapp. De Kinkerbuurt is vernoemd naar de belangrijkste straat, de Kinkerstraat. En deze straat is weer vernoemd naar Johann. Toen in 1937 de Kinkerbrug aan het einde van de Kinkerstraat in gebruik werd genomen, werd dit gevierd met de Wink. Want in de periode dat de Kinkerbuurt ontstond, ontstond ook iets anders. Iets dat voor ons heel normaal is, winkelen. Winkels kregen voor het eerst grote etalages waar de spullen werden uitgestald. Die grote etalages waren mogelijk doordat je het raam kon overspannen met een grote stalen balk. Waar er vroeger dus kleine ramen met een enkel uitgestald artikel, nu kon je langs de grote etalages lopen en. Zonder per se iets te kopen. Windowshoppen dus. Het winkelen nam in het begin van de vorige eeuw een enorme vlucht. Mensen kregen steeds meer geld te besteden en de Kinkerstraat ontwikkelde zich al snel tot de belangrijkste. De mooiste en duurste winkels van Amsterdam stonden echter niet in de Kinkestraat, maar in het centrum. Dit waren de warenhuizen en modepaleizen. En in het begin waren deze winkels uitsluitend voor de rijken. Denk hierbij aan Maison de Bonetterie aan de Kalverstraat of de Bijenkorf aan de Dam. De eigenaars. En om die reden opende men in 1926 de HEMA. Ofwel Hollandse Eenheidsprijzenmaatschappij. Bij de eerste HEMA's was niets duurder dan 50 cent. En de HEMA was een groot succes. En op 4 augustus 1928 opende de HEMA het eerste filiaal aan de Kinkerstraat. TV Gelderland 2021

Themes

1. Economie
2. Innovatie
3. Stedenbouw
4. Geschiedenis
5. Architectuur

Visualisation of result in (temporary) front-end

The screenshot displays a web interface with a search bar at the top containing the text 'Haarlemmerstraat' and a blue 'Search' button. Below the search bar, there are two expandable result sections. The first section, titled 'Amsterdamse doofpot', shows the following information: Entity: Haarlemmerstraat; Sentence: Vanuit het centrum loopt de menigte over de Nieuwe Dijk de Haarlemmerstraat in.; Timestamps: 59.18:127.56; Audio: (with a video player interface showing 0:00 / 0:00); and Adalink: <https://adalink.nl/geo/street/haarlemmerstraat/1629>. The second section, titled 'Bakkerij Jongejans', shows: Entity: Haarlemmerstraat; Sentence: Dit pand aan de Haarlemmerstraat is al sinds 1860 een bakkerij.; Timestamps: 32.32:34.5; Audio: (with a video player interface showing 0:00 / 0:00); and Adalink: <https://adalink.nl/geo/street/haarlemmerstraat/1629>.

30 second
video chunks

| Summary

Step 1: Speech-to-Text (**word level**)



Step2: Named Entity Recognition (**sentence level**)



Step 3: LLM based theme / concept extraction (**paragraph level**)



Using local models: on laptop or on-premise SURF AI-hub



Conclusion

Combining *existing* technologies and adding new value by

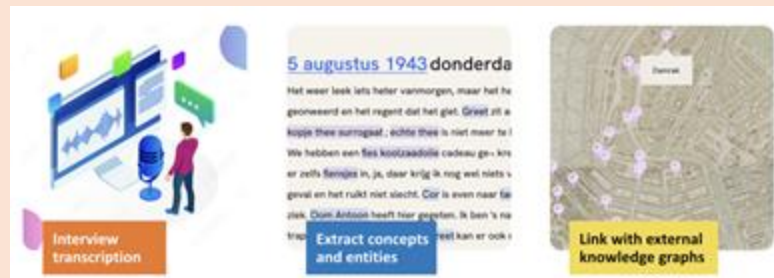
- An automated end-to-end solution

specific choices for this case:

entities, linked external resources, LLM (size depending on infrastructure)

- Context is important for LLMs -> short sentences are harder to interpret
- Modular: re-use of elements

New solution for this audience (data curators, conservation researchers)



Thanks to:

**Simone van Bruggen
(SURF)**

Design and testing of
the annotation pipeline

**Boudewijn Koopmans,
Ingeborg Verheul, Leon
van Wissen (UvA)**

Project team
Meaningful Memories



Code available at: <https://github.com/SURF-ML/meaningful-memories>



SURF

| Example prompt for LLM based theme search

```
{  
    "role": "system",  
    "content": "You are an assistant helping with finding relevant themes and  
concepts in a piece of Dutch text. Focus on larger and more abstract themes. Always  
reply in Dutch. Return the list of concepts only, do not explain yourself or summarize  
the text.",  
},  
{  
    "role": "user",  
    "content": "Welke thema's, concepten komen hier voor? Geef alleen de lijst  
met concepten in korte keywords zonder uitleg, gescheiden door een komma.",  
},
```


| Importance of context for LLMs

En er komen echt uit alle delen van de wereld schepen en mensen naar de stad.

En juist daarom wordt **Amsterdam** ook een belangrijk centrum voor informatie.

Concreet moet je toch vooral denken aan de **beurs** en aan de prijscuranten die gemaakt worden... op basis van de prijzen die



| Quality

No gold standard: no ground truth annotations for this dataset

WhisperX

Result accepted *as-is* (PoC)

Named Entity Recognition and LLM-based theme/concept extraction

How to estimate quality?

Probability per entity -> only accept if above threshold

Is probability a good measure of quality?

|Value for SURF, the Dutch IT cooperative for education and research

Potential interest for all researchers using interviews/spoken narratives

Including other fields (e.g. psychology, health domain)

Reference for future (Dutch) speech projects

Related to running projects

HOSAN (*Hoogwaardige Spraakherkenning voor het Nederlands*)

GPT-NL

