



Universität
Marburg

Online-Werkstattgespräch

Digitale Projekte eigenständig umsetzen

Felix Tacke

felix.tacke@uni-marburg.de

Veranstaltet von der AG Digitale Romanistik in Kooperation mit Text+

Worum geht es heute?

- Was man mittels **KI-Coding** alles tun kann
 - Wie setzt man **digitale Projekte** um?
 - Wie bastelt man sich **digitale Tools**?
 - Wie bastelt man sich **digitale Pipelines**?

Warum ich?




Christoph Niemann, „Prompt“, 2023

Angepasstes Open Source-Template
Mkdocs/Zensical-Theme
Text (.md, .html)
Multimedia

„Von Null gebastelt“

Große Webapp
Multimedia
Datenbanken
Search Engine

v0 (2023-2025)
v1 2026



CO.RA.PAN

A large-scale corpus of contemporary broadcast Spanish from nearly all Spanish-speaking countries. The web app integrates transcript navigation, metadata filters, and synchronized audio playback.


[Learn more →](#)



Linguistik im Spanischunterricht

A digital textbook for Spanish language teaching that connects linguistic expertise, multimedia materials, and modular didactic design in an openly accessible web publication.

[Learn more →](#)



Games.Hispanistica

A collection of gamified, web-based learning modules that translate established linguistic content into interactive quiz and game formats for analysis, comparison, and feedback-driven exploration.


[Learn more →](#)

Angepasst von **corapan-v1**

Quiz-Webapp
Datenbanken
Interaktive Games

Angepasst von **corapan-v0**


Kleine Webapp
Webtools



MAR.ELE

A pronunciation corpus capturing spoken Spanish from learners across proficiency levels and linguistic backgrounds. The platform combines recordings with transparent annotation workflows and speaker metadata.

[Learn more →](#)



Pronunciation Matters

Upcoming multilingual platform for learner pronunciation corpora. Extends MAR.ELE to English, French, Spanish, and German with a modern research and teaching interface.

[Learn more →](#)

Weiterentwickelt ausgehend von **corapan-v1**

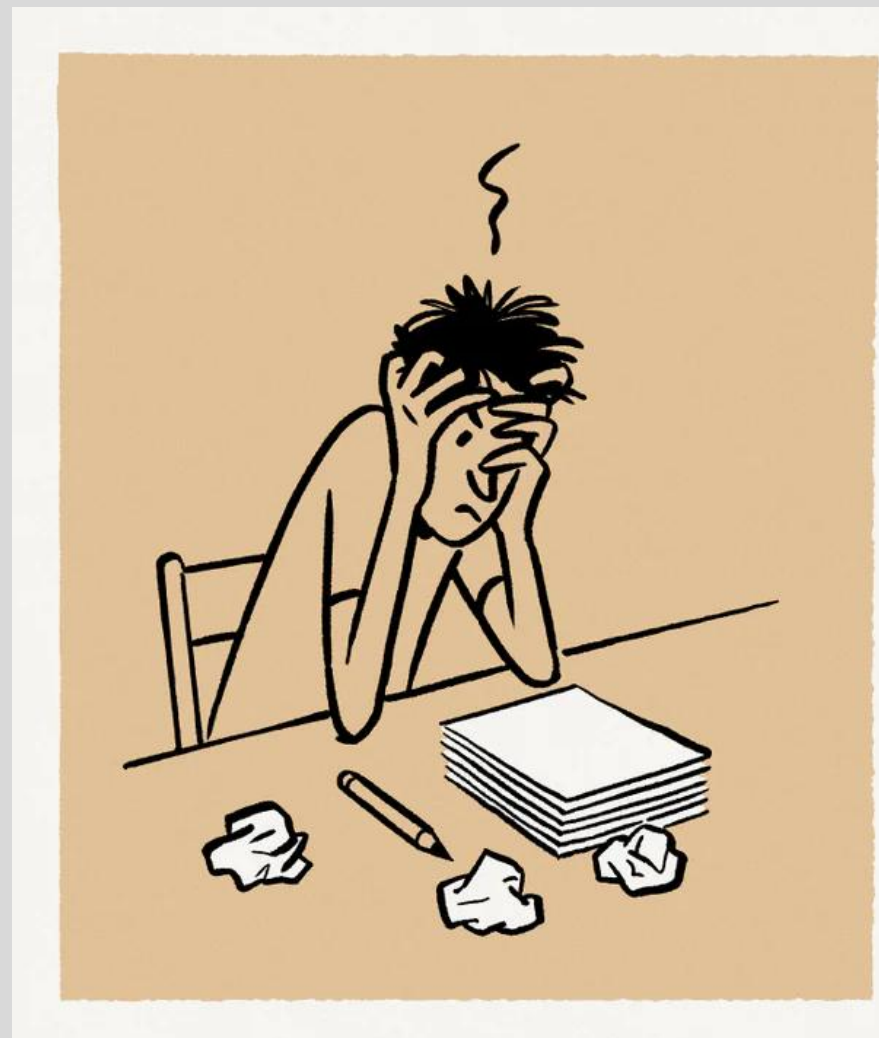
Webapp für
Forschung/ Lehre/
Schule
Datenbanken
Multimedia
Webtools

Immer dabei:
Verarbeitung von
(sprachlichen) Daten

- Systematisierung
- Anpassung
- Anreicherung
- Umwandlung

„Once Upon a Time“

... als es noch keine LLMs gab.



Christoph Niemann, „Prompt“, 2023

– Situation

- allgemein wenige Expert:innen für **Digital Humanities** in den Geisteswissenschaften
- wenige Institute/Fakultäten mit fester **IT-Abteilung**
- Spezialist:innen meist nur temporär über **Drittmittel**

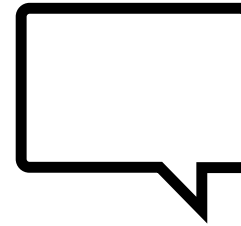
„Do It Yourself“ dank LLMs: (2022), 2023, 2024



Christoph Niemann, „Prompt“, 2023

„Do It Yourself“

– seit Ende 2022: **LLM als Programmierer:in**



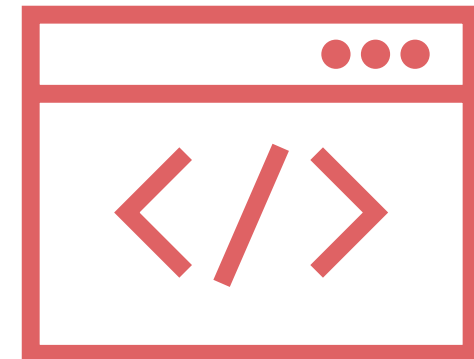
– neu: Programmieren mittels ‚**natürlicher Sprache**‘

1. Wir beschreiben in unserer Sprache, was wir erreichen möchten

2. LLMs verstehen, was wir wollen

& übersetzen es in Programmiersprachen (z.B. Python, Javascript)

& geben Anleitung „für Dummies“



„Do It Yourself“: Vibe Coding

Vibe Coding (1):

Ein intuitiver Programmieransatz, bei dem Entwickler Entscheidungen im Codefluss nach Gefühl, Kreativität und situativem Kontext statt nach festen Regeln treffen.

Vibe Coding (2):

Ein durch LLMs ermöglichter Programmierstil, bei dem Software in natürlicher Sprache beschrieben und automatisch in funktionierenden Code übersetzt wird.



Quelle: „Vibes won't cut it — Chris Kelly, Augment Code“, Youtube, 03.08.2025

„Do It Yourself“



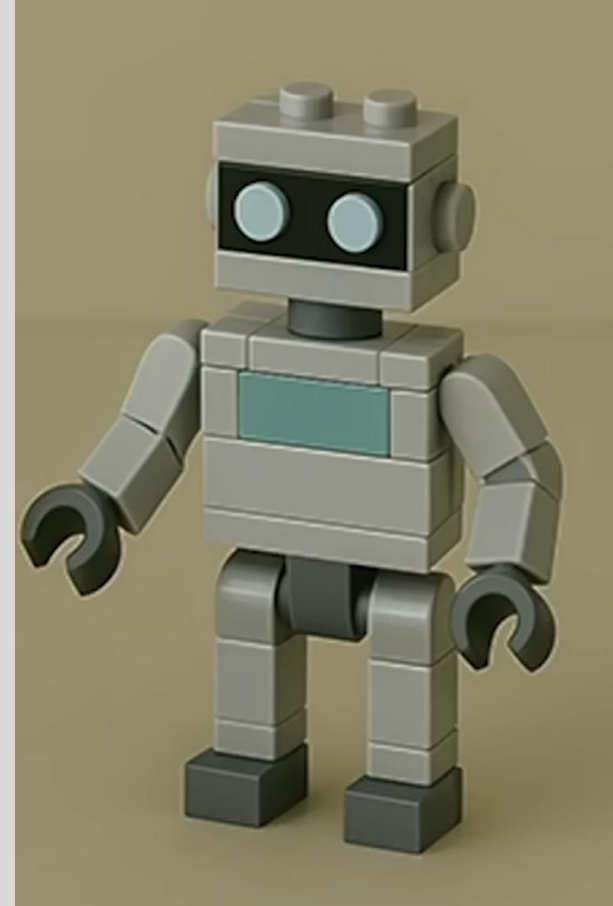
– Kommunikation mit LLM-Chatbot

- Beratung
- Code / Codeschnipsel
- Code-Überarbeitung

– *heuristisches* Vorgehen

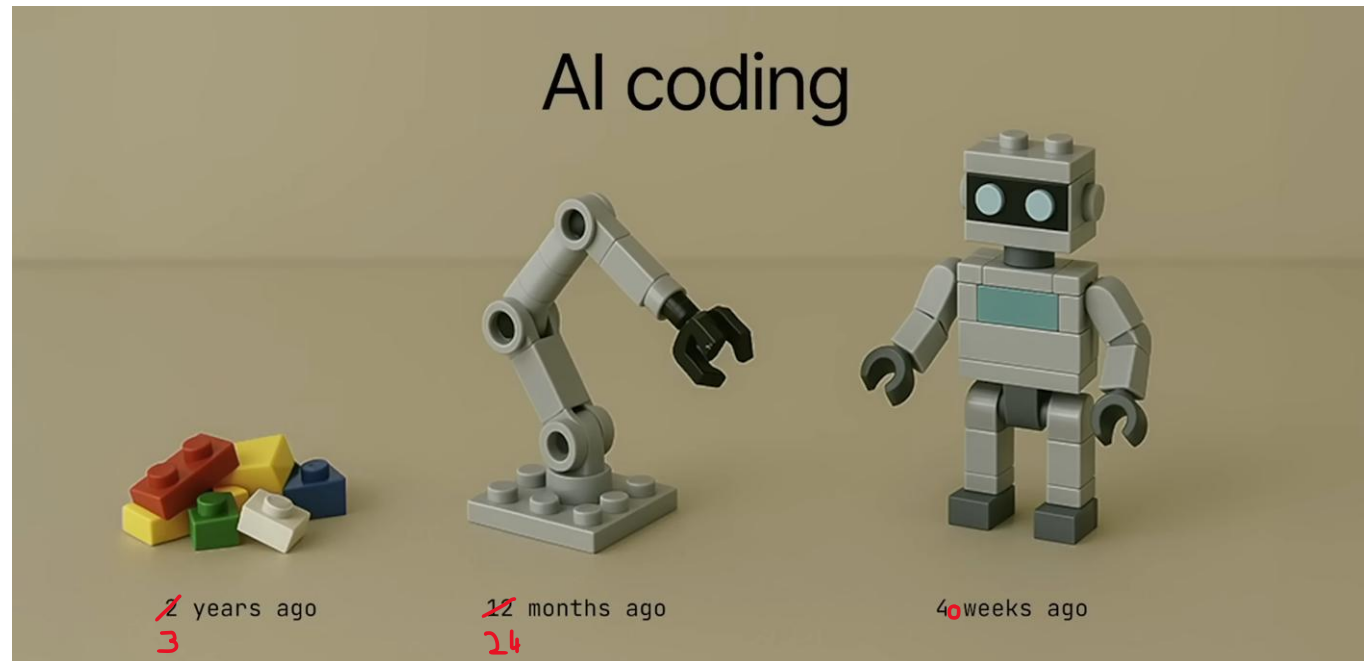
- LLM liefert Code im Chat → *Copy & Paste*
- ich teste Code, gebe Rückmeldung im Chat
 - kopiere Fehlermeldungen
 - erläutere Probleme
- LLM analysiert
 - liefert überarbeiteten Code / Codeschnipsel im Chat
→ *Copy & Paste*
- ...

„Let It Do It For You“: 2025ff.



Quelle: „Vibes won't cut it — Chris Kelly, Augment Code“, Youtube, 03.08.2025

„Let It Do It For You“



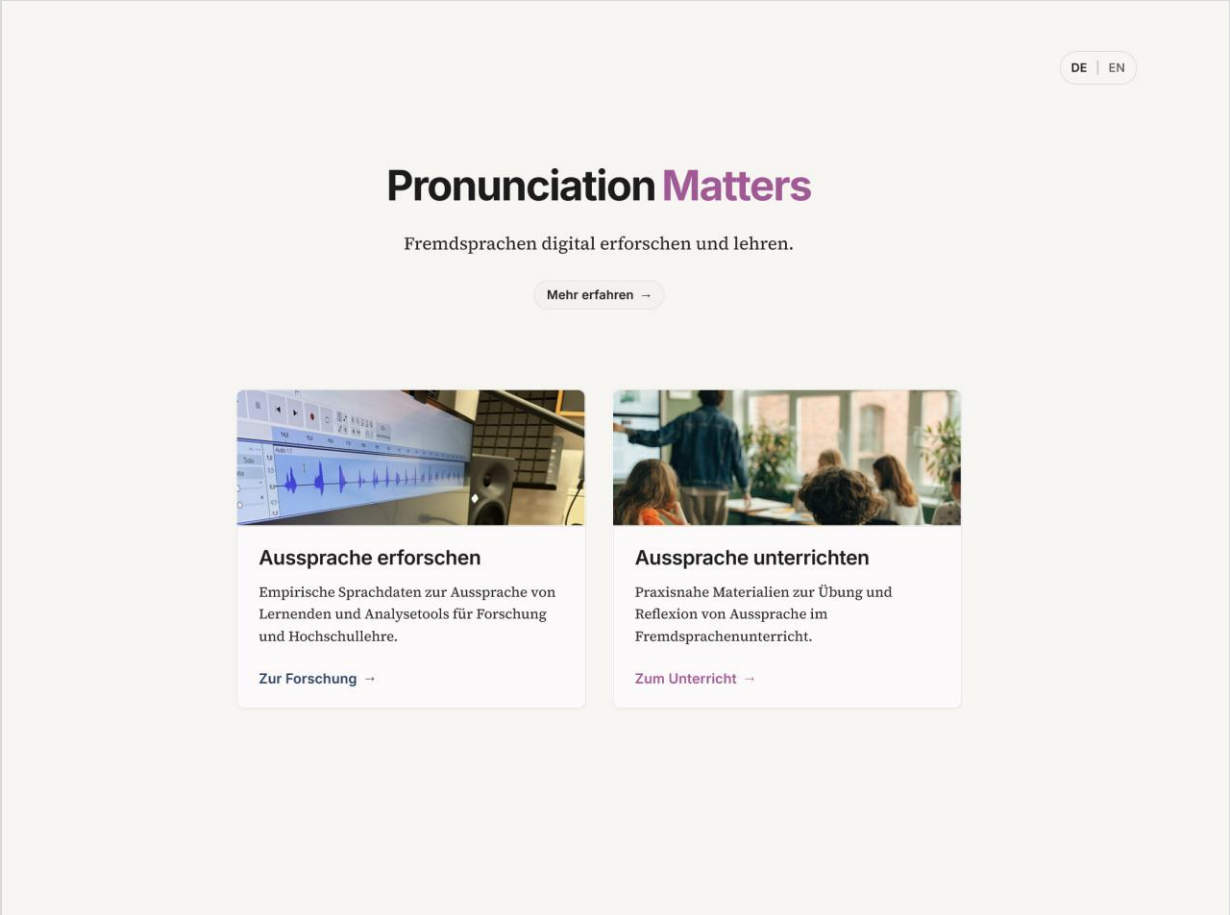
Quelle: „Vibes won't cut it — Chris Kelly, Augment Code“,
Youtube, 03.08.2025

„Let It Do It For You“

- Was bedeutet das?
 - **LLM-Agent** statt LLM-Chatbot
 - ist technisch direkt eingebunden (z.B. in die lokale Programmier-Software)
 - kann selber **agieren**
 - Projektdateien analysieren
 - Files erstellen
 - Änderungen umsetzen
 - Tests durchführen, Screenshots machen und analysieren, Code verbessern, etc.
 - Dokumentieren



Aktuelles Beispiel




The screenshot shows the homepage of the 'Pronunciation Matters' website. At the top right, there are language selection buttons for 'DE' and 'EN'. The main heading is 'Pronunciation Matters' in a purple and black font, followed by the tagline 'Fremdsprachen digital erforschen und lehren.' Below this is a button labeled 'Mehr erfahren →'. The page features two main content cards. The left card, titled 'Aussprache erforschen', includes a thumbnail image of a computer screen displaying a spectrogram and a microphone, and a description of empirical language data for research. The right card, titled 'Aussprache unterrichten', includes a thumbnail image of a teacher in a classroom and a description of practical materials for teaching. Both cards have corresponding action buttons: 'Zur Forschung →' and 'Zum Unterricht →'.

DE | EN

Pronunciation Matters

Fremdsprachen digital erforschen und lehren.


[Mehr erfahren →](#)



Aussprache erforschen

Empirische Sprachdaten zur Aussprache von Lernenden und Analysetools für Forschung und Hochschullehre.

[Zur Forschung →](#)

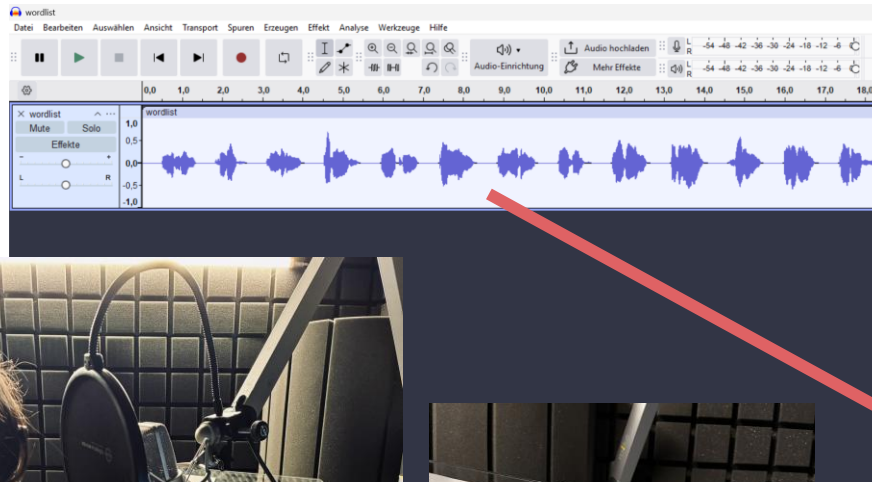


Aussprache unterrichten

Praxisnahe Materialien zur Übung und Reflexion von Aussprache im Fremdsprachenunterricht.

[Zum Unterricht →](#)

Daten strukturieren: Von A nach B kommen



ES-L-0005-2026-S01 Primär

Lernende **BT** L1 DE

Person-ID: ES-L-0005
Aufnahmedatum: 2026-02-23
Geschlecht: männlich
Sprachaufenthalte: Keine erfassten Sprachaufenthalte
Explorator:in: Marlon Merte

Profil + Vergleich

Wortliste **Text** Interview Set wählen i Alle Items

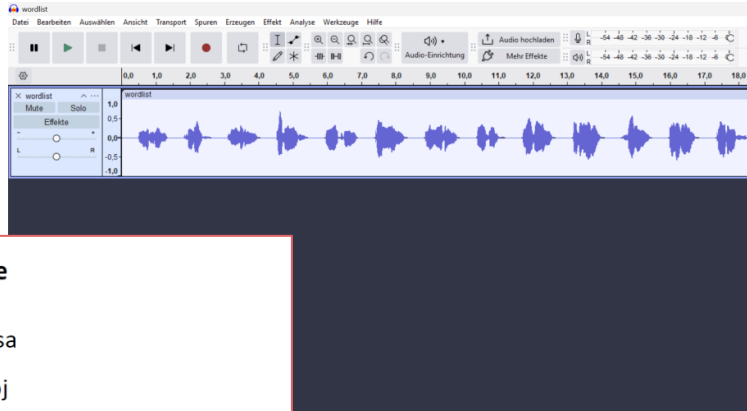
Wiedergabe
0:00 / 2:19 Lautstärke 100% Geschwindigkeit 1.00x

Wortliste
92 Items

1	mesa	0:00-0:01	↓
2	reloj	0:01-0:02	↓
3	viuda	0:03-0:03	↓
4	tabúes	0:04-0:05	↓
5	neutro	0:05-0:06	↓

Daten strukturieren: Von A nach B kommen

1



Wortliste

1. mesa
2. reloj
3. viuda
4. tabúes
5. neutro

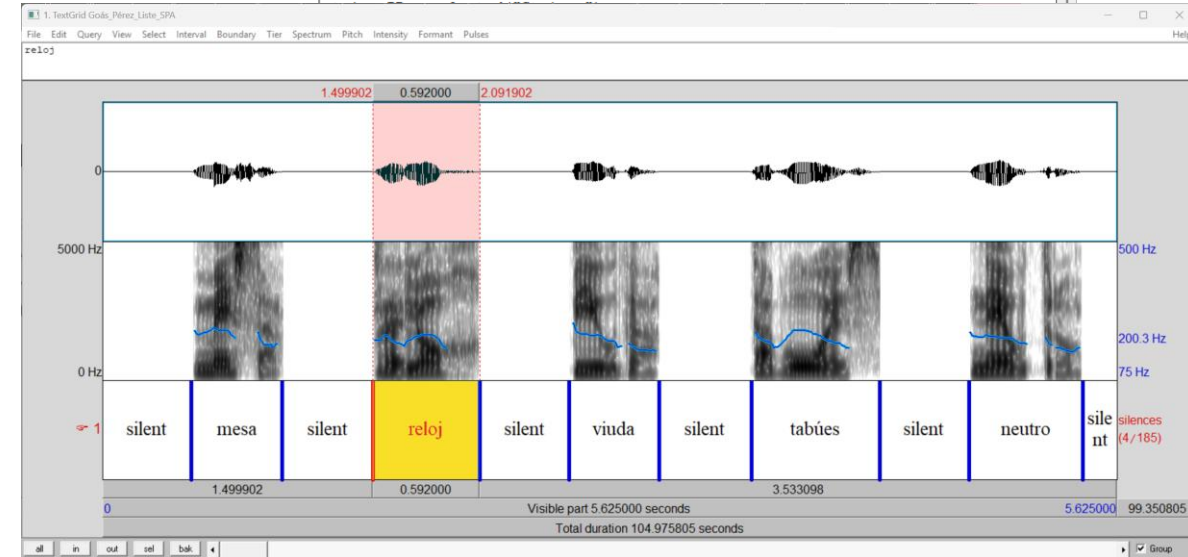
2

```
Script "C:/dev/promat-data/scripts/label_wortliste_from_sounding.praat"
File Edit Search Convert Font Run Help
# =====
# EINSTELLUNGEN
# =====
wordListFile$ = "C:/dev/promat-data/pipeline_list/raw_list/spanish_wordlist.txt"
textGridFolder$ = "C:/dev/promat-data/pipeline_list/01_praat_annotation/"
tierNumber = 1

# =====
# PRUEFEN: EIN TEXTGRID AUSGEWAHLT
# =====
textgridID = selected("TextGrid")
if textgridID = 0
  exitScript: "Bitte genau ein TextGrid im Objects-Fenster auswählen."
endif

selectObject: textgridID
textgridName$ = selected$("TextGrid")

# =====
# WORTLISTE LADEN
# =====
Read Strings from raw text file: wordListFile$
```



```
# =====
# LABELS SETZEN
# =====
wordIndex = 1
for i from 1 to numIntervals
  selectObject: textgridID
  label$ = Get label of interval: tierNumber, i

  if label$ = "sounding"
    selectObject: stringsID
    word$ = Get string: wordIndex

    selectObject: textgridID
```

Daten strukturieren: Von A nach B kommen

3

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 139.21478458049887
tiers? <exists>
size = 1
item []:
  item [1]:
    class = "IntervalTier"
    name = "silences"
    xmin = 0
    xmax = 139.21478458049887
    intervals: size = 185
    intervals [1]:
      xmin = 0
      xmax = 0.48339229024942837
      text = "silent"
    intervals [2]:
      xmin = 0.48339229024942837
      xmax = 1.3473922902494284
      text = "mesa"
    intervals [3]:
      xmin = 1.3473922902494284
      xmax = 1.8193922902494284
      text = "silent"
    intervals [4]:
      xmin = 1.8193922902494284
      xmax = 2.6113922902494284
      text = "reloj"
    intervals [5]:
      xmin = 2.6113922902494284
      xmax = 3.0913922902494284
      text = "silent"
    intervals [6]:
      xmin = 3.0913922902494284
      xmax = 3.9873922902494283
      text = "viuda"
    intervals [7]:
      xmin = 3.9873922902494283
      xmax = 4.523392290249429
      text = "silent"
    intervals [8]:
      xmin = 4.523392290249429
      xmax = 5.419392290249428
      text = "tabúes"
```

4

```
{
  "session_id": "ES-L-0005-2026-S01",
  "person_id": "ES-L-0005",
  "task": "wordlist",
  "audio": {
    "full_mp3": "derived/wordlist.mp3"
  },
  "items": [
    {
      "item_id": "wl_001",
      "item_number": "1",
      "text": "mesa",
      "start_ms": 483,
      "end_ms": 1347,
      "split_mp3": "items/wordlist/wl_001.mp3"
    },
    {
      "item_id": "wl_002",
      "item_number": "2",
      "text": "reloj",
      "start_ms": 1819,
      "end_ms": 2611,
      "split_mp3": "items/wordlist/wl_002.mp3"
    },
    {
      "item_id": "wl_003",
      "item_number": "3",
      "text": "viuda",
      "start_ms": 3091,
      "end_ms": 3987,
      "split_mp3": "items/wordlist/wl_003.mp3"
    },
    {
      "item_id": "wl_004",
      "item_number": "4",
      "text": "tabúes",
      "start_ms": 4523,
      "end_ms": 5419,
      "split_mp3": "items/wordlist/wl_004.mp3"
    },
    {
      "item_id": "wl_005",
      "item_number": "5",
      "text": "neutro",
      "start_ms": 5947,
      "end_ms": 6883,
      "split_mp3": "items/wordlist/wl_005.mp3"
    }
  ]
}
```



person_id	speaker_type	I1	I1_additional	mother_I1	father_I1	additional_languages	gender
ES-L-0001	learner	DE		DE	DE	EN; SV	female
ES-N-0001	native_speaker	ES		unknown	unknown		male
ES-N-0002	native_speaker	ES		unknown	unknown		female
ES-L-0002	learner	DE		DE	DE	EN; FR; AR	female
ES-L-0003	learner	DE		DE	DE	EN; IT	female
ES-L-0004	learner	DE		DE	DE	FR; EN	female
ES-L-0005	learner	DE		DE	DE	EN	male
ES-L-0006	learner	DE		DE	DE	EN	male
ES-L-0007	learner	DE		DE	DE	EN	male
ES-L-0008	learner	DE		DE	DE	EN; NL	female
ES-L-0009	learner	DE		DE	DE	EN; FR; IT	female
EN-L-0001	learner	DE		DE	DE	ES; SV	female
EN-L-0002	learner	DE		DE	DE	ES; FR; AR	female
EN-L-0003	learner	ES	DE	ES	DE	ES; FR	male
EN-L-0004	learner	DE		DE	DE	ES	male
EN-L-0005	learner	DE		DE	DE	IT	female
EN-L-0006	learner	ES	DE	ES	DE	FR	female
EN-L-0007	learner	ES	DE	ES	DE	FR	female
EN-L-0008	learner	DE		DE	DE	ES; NL	female

Was tut das Intake-Script (noch alles)?

5

```

Batch-Eingang
processed/*_wordlist_processed.wav
processed/*_wordlist_processed.TextGrid
raw/*_wordlist_raw.wav optional
├─┬─
│ │
│ │ import/organize_batch_working_tree.py
│ │ │
│ │ │ erzeugt:
│ │ │ │ working/{person_id}/wordlist/source/wordlist.wav
│ │ │ │ working/{person_id}/wordlist/alignment/wordlist.TextGrid
│ │ │
│ │ │
│ │ │ import_batch_to_production.py --sync-tasks
│ │ │ │
│ │ │ │ liest Workbook aus intake_data/*.xlsx
│ │ │ │ leitet session_id ab
│ │ │ │ schreibt/aktualisiert DB-Metadaten
│ │ │ │ archiviert raw/wordlist.wav, falls vorhanden
│ │ │ │ kopiert Working-Dateien nach data/sessions/...
│ │ │ │
│ │ │ │
│ │ │ │ intern: produce_wordlist_artifacts.py
│ │ │ │ │
│ │ │ │ │ erzeugt derived/wordlist.mp3
│ │ │ │ │ erzeugt alignment/wordlist.json
│ │ │ │ │ schneidet items/wordlist/*.mp3
│ │ │ │ │
│ │ │ │ │
│ │ │ │ │ metadata.json wird geschrieben/aktualisiert

```

6

```

├─┬─ sessions
│ │
│ │ > english
│ │
│ │ > spanish
│ │ │
│ │ │ > ES-L-0001-2026-S01
│ │ │ > ES-L-0002-2026-S01
│ │ │ > ES-L-0003-2026-S01
│ │ │ > ES-L-0004-2026-S01
│ │ │ > ES-L-0005-2026-S01
│ │ │ │
│ │ │ │ > alignment
│ │ │ │ > derived
│ │ │ │ > items
│ │ │ │ > raw
│ │ │ │ > source
│ │ │ │ {} metadata.json
│ │ │ │
│ │ │ │ > ES-L-0006-2026-S01
│ │ │ │ > ES-L-0007-2026-S01
│ │ │ │ > ES-L-0008-2026-S01
│ │ │ │ > ES-L-0009-2026-S01
│ │ │ │ > ES-N-0001-2026-S01
│ │ │ │ > ES-N-0002-2026-S01

```

```

├─┬─ ES-L-0005-2026-S01
│ │
│ │ > alignment
│ │ │
│ │ │ {} interview.json
│ │ │ {} text.json
│ │ │ ≡ text.TextGrid
│ │ │ {} wordlist.json
│ │ │ ≡ wordlist.TextGrid
│ │
│ │ > derived
│ │ │
│ │ │ 🔊 interview.mp3
│ │ │ 🔊 text.mp3
│ │ │ 🔊 wordlist.mp3
│ │
│ │ > items
│ │ │
│ │ │ > text
│ │ │
│ │ │ > wordlist
│ │ │ │
│ │ │ │ 🔊 wl_001.mp3
│ │ │ │ 🔊 wl_002.mp3
│ │ │ │ 🔊 wl_003.mp3
│ │ │ │ 🔊 wl_004.mp3
│ │ │ │ 🔊 wl_005.mp3
│ │ │ │ 🔊 wl_006.mp3
│ │ │ │ 🔊 wl_007.mp3
│ │ │ │ 🔊 wl_008.mp3

```

Wie gehe ich da ran?

– Teil 1: Konzept erarbeiten

- Was habe ich?
(hier: Audiofile, Textfile)
- Was brauche ich?
(hier: Files mit Text + Zeitstempeln)
- Wofür brauche ich das?
(hier: Webapp-Player)
- Was muss ich bedenken?
(hier: Verknüpfung Metadaten, Daten, Webapp)

Arbeite mit einem LLM-Chatbot! (z.B. Claude, ChatGPT)

1. Kontext geben

- gib **Material** in den Chat (Textfiles, Ausschnitte, Screenshots)
- **beschreibe**, was Du willst; lass dir Vorschläge machen; **frage nach**, wenn genannte Optionen unklar sind
- kläre offene Fragen: „wäre JSON oder CSV oder ein anderes Format besser für meine Daten?“
- **entscheide** dich für Optionen („ich will Variante B“)
- mache **Anmerkungen**: „besonders wichtig ist mir, dass...“, „X passt, aber ich Y muss auch möglich sein“; „Option C ist für mich quatsch, weil...“

2. Überblick behalten

- bitte gelegentlich um **Zusammenfassungen**: „gib mir einen strukturierten Überblick über alles, was wir besprochen und entschieden haben“, „welche Fragen sind jetzt noch zu klären?“
- korrigiere Missverständnisse, präzisiere Anforderungen

3. Masterplan erstellen

- Lasse den Plan ausformulieren und **als strukturiertes Markdown-Dokument (.md)** ausgeben.

Beispiel Masterplan

https://notes.hispanistica.com/promat_audio

https://notes.hispanistica.com/promat_json

Wie gehe ich da ran?

– Teil 2: Konzept umsetzen

- Masterplan in die Tat umsetzen
(hier: Dateien physisch ordnen, Scripte erstellen, Scripte Daten umstrukturieren lassen)

Arbeite mit

a) VSC (Visual Studio Code) + KI-Agenten

b) LLM-Chatbot

1. VSC: Workspace

- lege einen leeren **Projektordner** auf deinem PC an und öffne ihn in VSC
- lege darin einen Unterordner /docs an und lege den Masterplan.md hinein

2. Chatbot (dein Planungsassistent)

- sage dem Chatbot, dass der Masterplan im Projektordner liegt und der KI-Agent darauf Zugriff hat; bitte um Prompt für die konkrete Umsetzung; Prompt soll dazu auffordern, die Ergebnisse des „Run“ in eine .md zu schreiben

3. VSC

- gib den Prompt in das Agenten-Chatfenster ein (ggf. „Approvals“ anpassen).
- gib die Ergebnis-md dem Chatbot zur Prüfung; er soll einen Prompt mit eventuell notwendigen Korrekturen erstellen bzw. einen Prompt für den nächsten Umsetzungsschritt

(...)

Visual Studio Code (VSC) + Github Education



- **Visual Studio Code** = Programmiersoftware (Microsoft, kostenlos); bietet Möglichkeit KI-Agenten zu integrieren
- **Github** = Online-Plattform für die Archivierung/Versionierung von Software
- **Github Education** = bietet kostenlosen Zugang zu **Github Copilot Pro**, worüber man ein monatliches Kontingent an KI-Zugriffen (verschiedene Anbieter/Modelle) erhält; deutlich günstiger als über API

Github Copilot Pro Zugang

1. GitHub-Account nutzen/erstellen.
2. Auf github.com/settings/education/benefits gehen.
3. Unter **GitHub Education** → „**Start an application**“.
4. Formular ausfüllen, Uni-Mail verwenden, Nachweis hochladen.
5. Nach Genehmigung Education-Portal nutzen.
 - 300 kostenlose „Premium-Anfragen“/Monat
 - weitere kosten per Guthaben \$0.04/Anfrage

Visual Studio Code (Microsoft, kostenlos)

geöffnete Datei

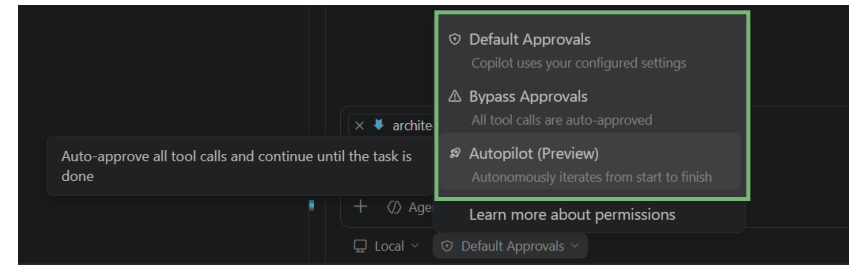
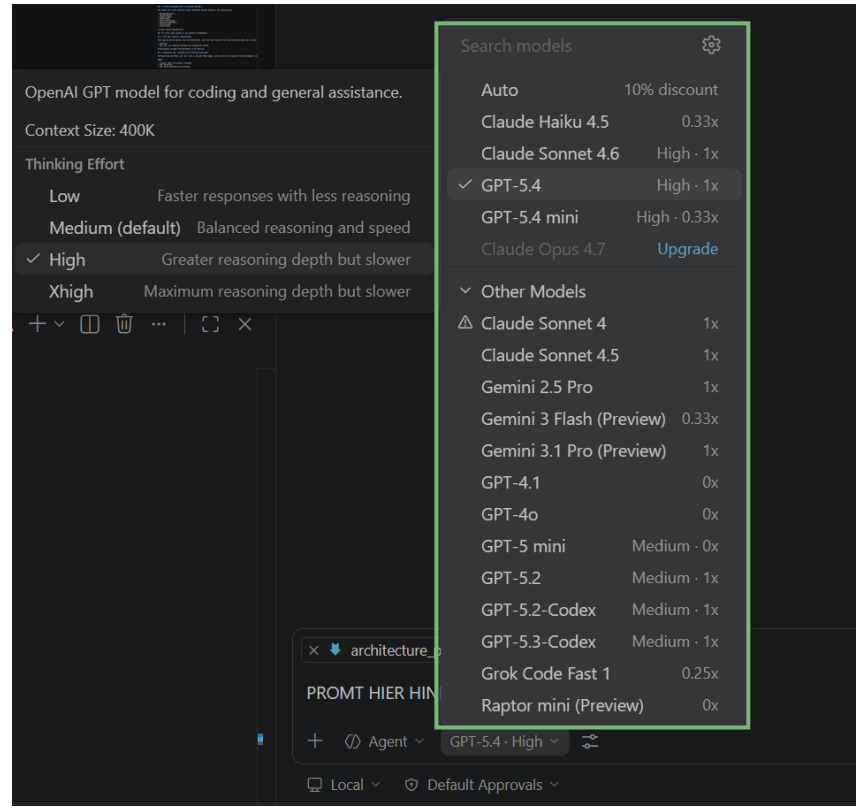
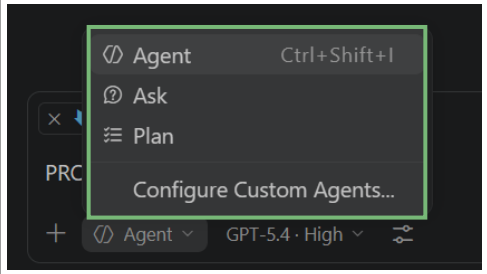
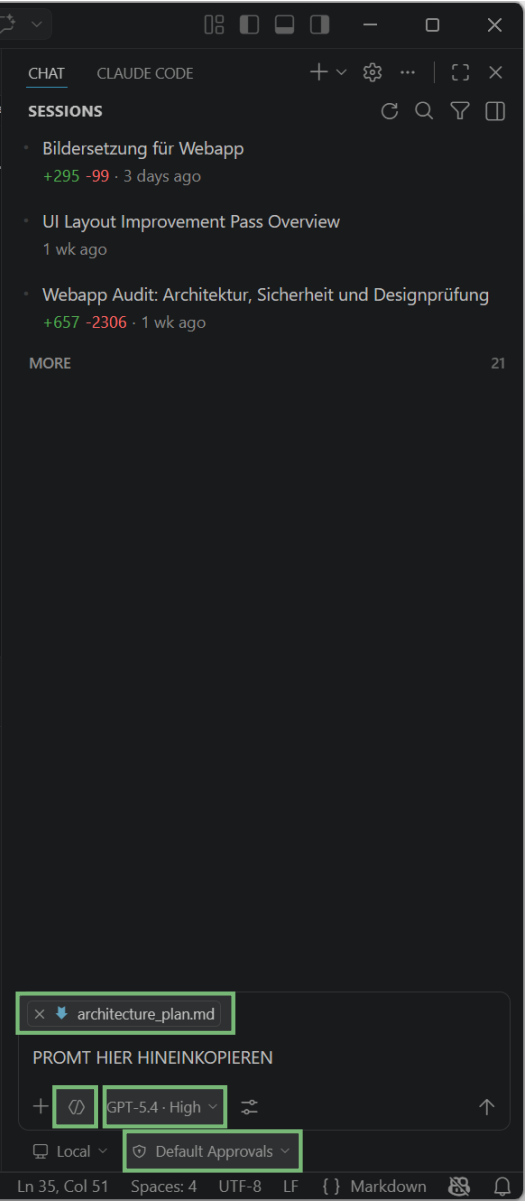
Projektordner/
Workspace
Ordner,
Files

The screenshot displays the Visual Studio Code interface with three main sections:

- Left Panel (Project Explorer):** Shows a file tree for a project named 'PROMAT'. The 'docs' folder is expanded, showing 'architecture_plan.md' selected.
- Center Panel (Editor):** Displays the content of 'architecture_plan.md'. The text includes a title 'tags: promat, webdesign, planung', a subtitle '# Architektur-Konsolidierung und Player-Optimierung', and a section '## Status und Zweck'. The main body contains a detailed project plan with phases 1 through 5, describing the consolidation of access, research capabilities, unified players, and set models. It also lists 'Ausgangslage' (Starting Point) and 'Ziele' (Goals).
- Right Panel (Chat):** Shows a chat interface with a 'SESSIONS' list. The current session is titled 'Bildersetzung für Webapp' and contains a message from 'KI-Agent' regarding 'Webapp Audit: Architektur, Sicherheit und Designprüfung'.

KI-Agent

Terminal (wo Befehle eingegeben werden)
[das kann heute alles der AGENT übernehmen!]



Wie gehe ich da ran?

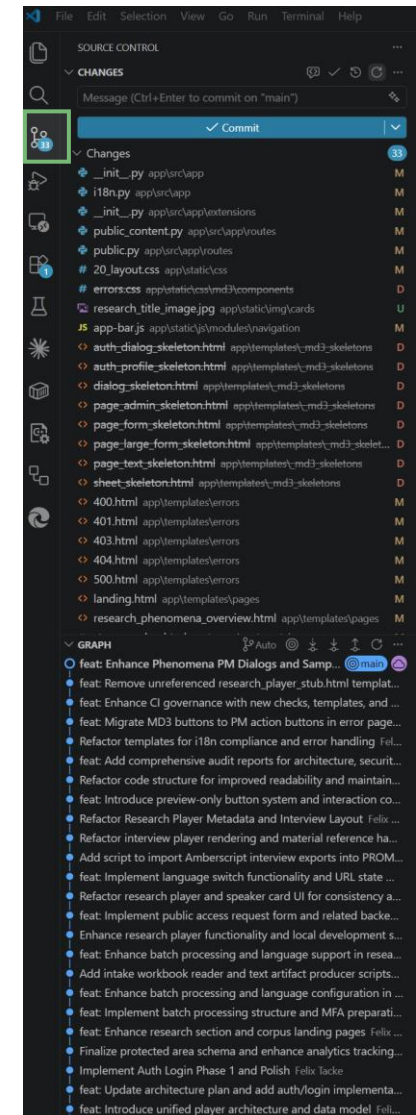
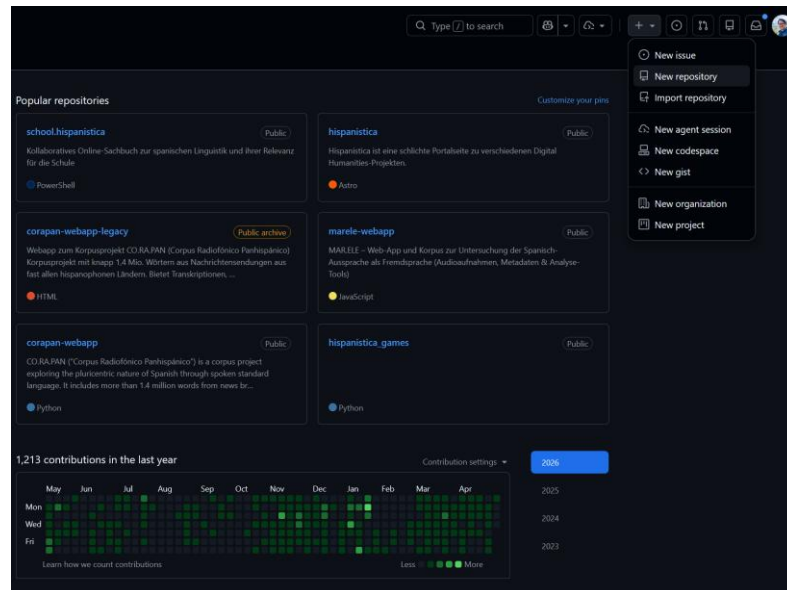


– Teil 3: Absicherung/Versionierung

- Nutze Git: GitHub (Microsoft) oder GitLab (deine Uni)

1. Github/Gitlab-Account erstellen
2. Repository für dein Projekt anlegen
3. VSC: Projektordner mit Repository verknüpfen

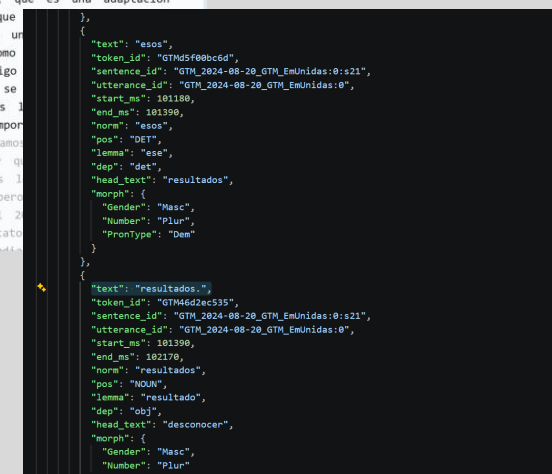
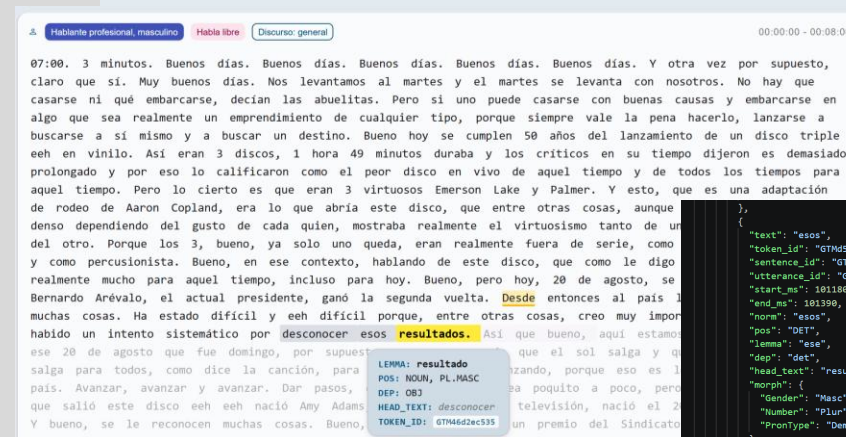
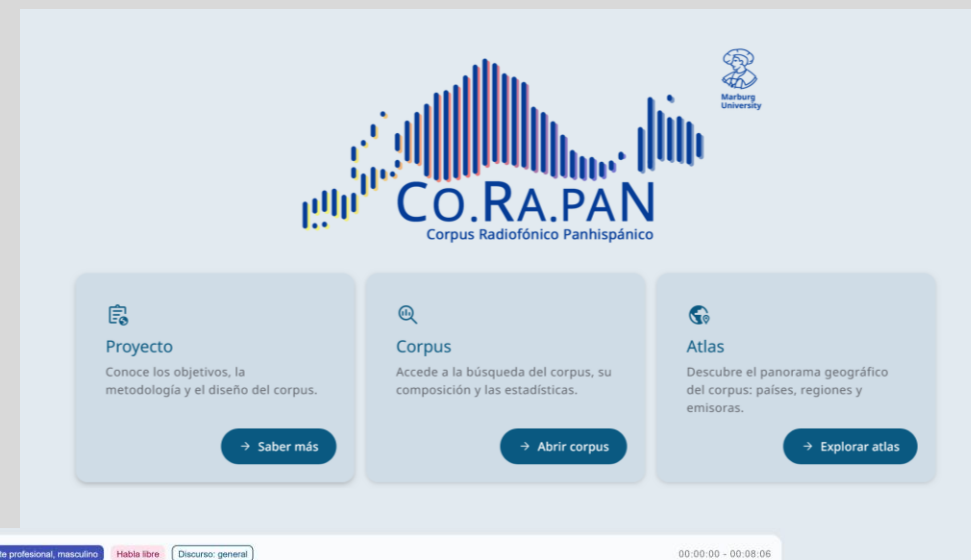
- frage deinen Chatbot, wie du VSC mit deinem Github/Gitlab-Account verknüpfst und wie du den Projektordner mit dem neuen Repository verbindest



Weiteres Beispiel

AI + NLP in

CO.RA.PAN



← 2024-08-20_GTM_EmUnidas.mp3

Hablante profesional, masculino Habla libre Discurso: general 00:00:00 - 00:08:06

07:00. 3 minutos. Buenos días. Buenos días. Buenos días. Buenos días. Buenos días. Y otra vez por supuesto, claro que sí. Muy buenos días. Nos levantamos al martes y el martes se levanta con nosotros. No hay que casarse ni qué embarcarse, decían las abuelitas. Pero si uno puede casarse con buenas causas y embarcarse en algo que sea realmente un emprendimiento de cualquier tipo, porque siempre vale la pena hacerlo, lanzarse a buscarse a sí mismo y a buscar un destino. Bueno hoy se cumplen 50 años del lanzamiento de un disco triple eeh en vinilo. Así eran 3 discos, 1 hora 49 minutos duraba y los críticos en su tiempo dijeron es demasiado prolongado y por eso lo calificaron como el peor disco en vivo de aquel tiempo y de todos los tiempos para aquel tiempo. Pero lo cierto es que eran 3 virtuosos Emerson Lake y Palmer. Y esto, que es una adaptación de rodeo de Aaron Copland, era lo que abría este disco, que entre otras cosas, aunque sea un poco o muy denso dependiendo del gusto de cada quien, mostraba realmente el virtuosismo tanto de uno como de otro como del otro. Porque los 3, bueno, ya solo uno queda, eran realmente fuera de serie, como pianista, como bajista y como percusionista. Bueno, en ese contexto, hablando de este disco, que como le digo () 3 discos, era realmente mucho para aquel tiempo, incluso para hoy. Bueno, pero hoy, 20 de agosto, se cumple un año de que

Bernardo Arévalo, el actual presidente, ganó la segunda vuelta. Desde entonces al país le ha tocado padecer muchas cosas. Ha estado difícil y eeh difícil porque, entre otras cosas, creo muy importante subrayarlo. Ha habido un intento sistemático por desconocer esos resultados. Así que bueno, aquí estamos. Un año después de

LEMMA: resultado
POS: NOUN, PL.MASC
DEP: OBJ
HEAD_TEXT: desconocer
TOKEN_ID: GTM46d2ec535

Bisheriges Ergebnis: automatisch annotierte Objekte/Sätze

```
},
{
  "text": "esos",
  "token_id": "GTMd5f00bc6d",
  "sentence_id": "GTM_2024-08-20_GTM_EmUnidas:0:s21",
  "utterance_id": "GTM_2024-08-20_GTM_EmUnidas:0",
  "start_ms": 101180,
  "end_ms": 101390,
  "norm": "esos",
  "pos": "DET",
  "lemma": "ese",
  "dep": "det",
  "head_text": "resultados",
  "morph": {
    "Gender": "Masc",
    "Number": "Plur",
    "PronType": "Dem"
  }
},
{
  "text": "resultados.",
  "token_id": "GTM46d2ec535",
  "sentence_id": "GTM_2024-08-20_GTM_EmUnidas:0:s21",
  "utterance_id": "GTM_2024-08-20_GTM_EmUnidas:0",
  "start_ms": 101390,
  "end_ms": 102170,
  "norm": "resultados",
  "pos": "NOUN",
  "lemma": "resultado",
  "dep": "obj",
  "head_text": "desconocer",
  "morph": {
    "Gender": "Masc",
    "Number": "Plur"
  }
}
```

– Bisher:

1. Transkription via Amberscript (KI-gestützt)
2. menschliche Korrektur von Inhalt, Wortlaut und Turn-/Sprecherzuordnung
3. **spaCy** mit **es_dep_news_trf** (Transformer-basiertes spanisches Modell für Morphologie, Lemmatisierung und Dependency Parsing)

– **Problem:** spaCy/es_dep_news_trf annotiert basierend auf den vorhandenen Satzgrenzen und der Interpunktion.

– Zukünftig:

- **vorgeschaltete KI (DeepSeek V4 Pro)** prüft Satzgrenzen und Interpunktion im Turn-Kontext
- **spaCy** erzeugt daraufhin zuverlässigere POS-, Lemma-, Dependency- und Morphologie-Annotation
- Kernidee: KI ersetzt spaCy nicht. Sie verbessert die Eingabe, damit spaCy auf besser strukturierten Turns annotiert

Die gesamte Pipeline soll über ein Script in VSC gestartet werden und automatisiert durchlaufen:

- DeepSeek wird über einen API-Schlüssel integriert (pro Token berechnet, aber sehr günstig)
 - spaCy läuft lokal (kostenlos)

AI + NLP in CO.RA.PAN

Analyse (nur Textresultat) und Vorschlag von DeepSeek V4 PRO: Probeanalyse im Chat

Gegebene Sequenz (Wörter fix, nur Interpunktion variabel):

text Kopieren Herunterladen

Desde entonces al país le ha tocado padecer muchas cosas. Ha estado difícil y eeh difícil porque, entre otras cosas, creo muy importante subrayarlo. Ha habido un intento sistemático por desconocer esos resultados.

Problem:
Die Punkt-Grenzen trennen eine syntaktische Einheit in drei Teile:

- *Ha estado difícil porque ha habido un intento sistemático...*
- *Dazwischen eingeschoben: entre otras cosas, creo muy importante subrayarlo*

Der zweite Satz endet mit *subrayarlo*, aber das Objekt zu *subrayar* fehlt – es steht erst im nächsten Satz (*Ha habido un intento...*). Umgekehrt ist der dritte Satz (*Ha habido un intento...*) ein Nebensatz ohne Hauptsatz.

Vorschlag für geänderte Interpunktion (Wörter unverändert):

text Kopieren Herunterladen

Desde entonces al país le ha tocado padecer muchas cosas; ha estado difícil y difícil, porque, entre otras cosas -creo muy importante subrayarlo-, ha habido un intento sistemático por desconocer esos resultados.

Original	Neu	Begründung
cosas.	cosas;	Semikolon statt Punkt – die beiden Hauptsätze hängen eng zusammen
difícil y eeh difícil porque,	difícil y difícil, porque,	Wegfall des Füllers „eeh“ ist keine Wortänderung (Füllpartikel, kein lexikalisches Wort); Komma vor „porque“
subrayarlo. Ha habido	subrayarlo-, ha habido	Punkt entfällt, stattdessen Gedankenstrich-Komma für den Parenthese-Abschluss

AI + NLP in CO.RA.PAN

← 2024-08-20_GTM_EmUnidas.mp3

Hablante profesional, masculino Habla libre Discurso: general 00:00:00 - 00:08:06

07:00. 3 minutos. Buenos días. Buenos días. Buenos días. Buenos días. Buenos días. Y otra vez por supuesto, claro que sí. Muy buenos días. Nos levantamos al martes y el martes se levanta con nosotros. No hay que casarse ni qué embarcarse, decían las abuelitas. Pero si uno puede casarse con buenas causas y embarcarse en algo que sea realmente un emprendimiento de cualquier tipo, porque siempre vale la pena hacerlo, lanzarse a buscarse a sí mismo y a buscar un destino. Bueno hoy se cumplen 50 años del lanzamiento de un disco triple eeh en vinilo. Así eran 3 discos, 1 hora 49 minutos duraba y los críticos en su tiempo dijeron es demasiado prolongado y por eso lo calificaron como el peor disco en vivo de aquel tiempo y de todos los tiempos para aquel tiempo. Pero lo cierto es que eran 3 virtuosos Emerson Lake y Palmer. Y esto, que es una adaptación de rodeo de Aaron Copland, era lo que abría este disco, que entre otras cosas, aunque sea un poco o muy denso dependiendo del gusto de cada quien, mostraba realmente el virtuosismo tanto de uno como de otro como del otro. Porque los 3, bueno, ya solo uno queda, eran realmente fuera de serie, como pianista, como bajista



DeepSeek V4 PRO:

text Kopieren Herunterladen

Original: ...eran 3 virtuosos Emerson Lake y Palmer. Y esto, que es una adaptación de rodeo de Aaron Copland, era lo que abría este disco, que...

Vorschlag: ...eran 3 virtuosos, Emerson, Lake y Palmer, y esto -que es una adaptación de rodeo de Aaron Copland- era lo que abría este disco, que...

1. Schritt: Masterplan erstellen

 VERÄNDERT VOR 26 MINUTEN  PRIVATE

Umsetzungsplan: CO.RA.PAN-Pipeline mit Preprocess-Migration, DeepSeek-QA und finaler spaCy-Annotation

0. Ziel

Die Pipeline soll so umgebaut werden, dass spanische Radio-Transkripte vor der finalen linguistischen Annotation strukturell bereinigt werden. DeepSeek soll dabei kontrolliert Satzgrenzen, Interpunktion, Eigennamen-Hinweise und bestimmte Korrekturvorschläge liefern. spaCy soll danach die finale Annotation neu erzeugen.

Der zentrale Grundsatz lautet:

```
Erst Transkriptstruktur und Interpunktion bereinigen,  
dann DeepSeek kontrolliert prüfen lassen,  
danach spaCy final annotieren.
```

Nicht:

```
spaCy zuerst,  
danach DeepSeek.
```

Der Grund: Im aktuellen Skript entstehen `sentence_id`, `utterance_id`, `pos`, `lemma`, `dep`, `head_text` und `morph` auf Basis der vorhandenen Satzteilung. Diese Satzteilung wird derzeit im Skript schlicht über Satzzeichen wie `.`, `?`, `!` vorgenommen. Wenn die Satzgrenzen vorher falsch sind, annotiert spaCy auf einer falschen Struktur.



Die bestehende README beschreibt Step 02 bereits als zentrale spaCy-Annotation mit Token-IDs, Satz-/Utterance-IDs, POS, Lemma, Dependency und Morphologie. Deshalb muss die DeepSeek-basierte Strukturkorrektur vor diesem Schritt stattfinden.

Umsetzungsplan: CO.RA.PAN-Pipeline mit Preprocess-...

1. Zwei Betriebsmodi
2. Neue Zielpipeline
3. Zentrale Schema-Änderung: Interpunktion aus text ...
4. Sonderfälle im Preprocess
5. Zahlenregel
6. Regionale QC: Voseo in Argentinien und Uruguay
7. DeepSeek: Grundprinzip
8. DeepSeek-Step 1: Interpunktion und Satzgrenzen
9. DeepSeek-Step 2: Entity-Hints
10. DeepSeek-Step 3: Lexical-QA
11. DeepSeek-Step 4: Voseo-QA
12. Übergabeformat an DeepSeek
13. Rekonstruktion für spaCy
14. Anpassung von 02_annotate_transcripts_v3.py
15. QA und Validierung
16. Teststrategie
17. Empfohlene neue Skripte
18. Minimales Ziel-JSON nach Umbau
19. Copilot-Agent-Aufgabenpaket
20. Endgültige Leitlinien
21. Kurzfassung der finalen Architektur

[Expand all](#)
[Back to top](#)
[Go to bottom](#)

Q & A

+ Stelle irgendeine Frage  



Universität
Marburg

Jetzt ihr!