FACULTEIT LETTEREN EN WIJSBEGEERTE

REEDS – AMSTERDAM
30.06.2023

TOWARDS A PARSED CORPUS OF SOUTHERN DUTCH DIALECTS

Anne-Sophie Ghyselen, Melissa Farasyn & Anne Breitbarth

UNIVERSITEIT GENT

Research Foundation Flanders
Opening new horizons

GCND

Gesproken	*Spoken*
Corpus	*Corpus*
(Zuidelijk-)**N**ederlandse	*(Southern-)Dutch*
Dialecten	*Dialects*

UNIVERSITEIT GENT

The Dutch dialect area

GCND

Gesproken	*Spoken*
Corpus	*Corpus*
(Zuidelijk-)**N**ederlandse	*(Southern-)Dutch*
Dialecten	*Dialects*

UNIVERSITEIT GENT

GCND

2017: pilot studies (smaller grants FWO/Provinces Zeeland, West Flanders and East Flanders)

Funding for 'Medium-scale research infrastructure'
Flemish Research Foundation (I010120N)
2019-end of May 2024

UNIVERSITEIT GENT

Research Foundation Flanders
Opening new horizons

PARTNERS

GHENT UNIVERSITY

Meertens Instituut

/instituut voor de Nederlandse taal/

rijksuniversiteit groningen

TODAY

1. Prehistory
2. Project goals
3. Workflow
4. Future plans

UNIVERSITEIT
GENT

FACULTEIT LETTEREN
EN WIJSBEGEERTE

PREHISTORY

UNIVERSITEIT
GENT

THE RECORDINGS

Prof. W. Pée
1903-1986

Prof. V. Vanacker
1921-1999

- 1960s-1970s
- Dialect recordings in as many locations as possible

UNIVERSITEIT GENT

SPEAKERS

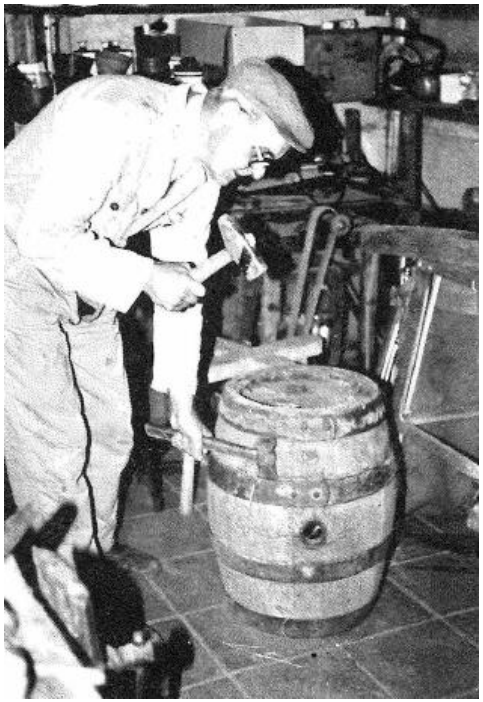
- Born around 1900 (oldest: °1871)
- Semi- and unskilled, many illiterate
- Non-mobile

UNIVERSITEIT GENT

CONTENTS: LIFE STORIES

- Technological innovations: electricity, cars, bikes…
- Traditional agriculture
- Old, vanishing crafts
- World wars
- Traditions, customs
- …

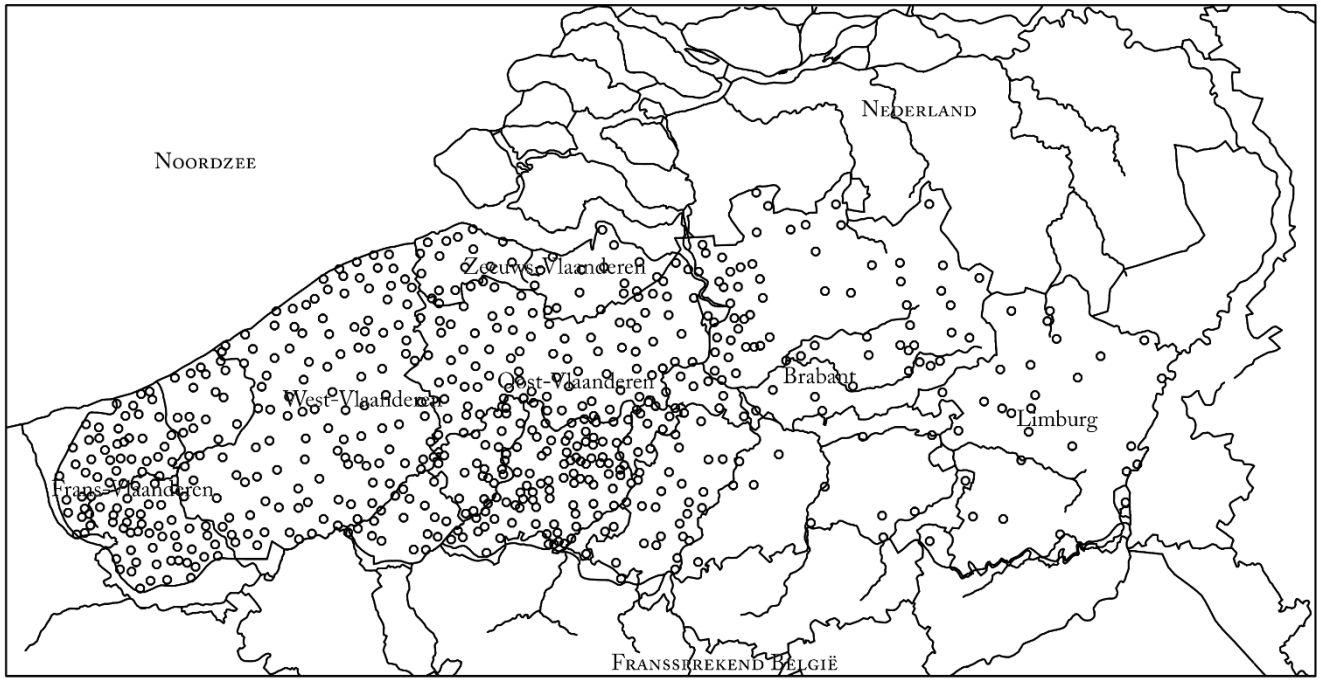
Bron: Erfgoedbank Waasland, Collectie Gemeentearchief Temse



Bron: Woordenboek van de Vlaamse Dialecten

UNIVERSITEIT GENT

RESULT



- 783 recordings
- 550 places
- ca. 700 hours
- Flanders, French Flanders and Zealand Flanders

UNIVERSITEIT GENT

VALUABLE DATA

- **Spontaneously spoken dialect data**: wealth of opportunities for syntactic research (cf. talk Anne Breitbarth 29.06.2023)
- **Heritage preservation**
  - **Document vanishing dialects**
  - ***Oral history***
  - **>** Document several types intangible cultural heritage

UNIVERSITEIT
GENT

BUT: MATERIAL: LONG UNDERUSED

UNIVERSITEIT GENT


Mail Lite
Mail Lite

RIJKSUNIVERSITEIT GENT

Academiejaar 1974 - 1975

ENKELE SYNTACTISCHE KENMERKEN VAN HET WICHELS DIALECT

DEEL I

Promotor:
Prof. Dr. V.F. VANACKER

Proefschrift voorgelegd aan de
Faculteit van de Letteren en Wijsbegeerte
voor het verkrijgen van de graad van
licentiaat in de Germaanse Filologie
door
Jacques VAN KEYMEULEN

Rijksuniversiteit te Gent
Faculteit Letteren en Wijsbegeerte
Academiejaar 1974 – 1975

ENKELE SYNTACTISCHE KENMERKEN VAN HET MALDEGEMS DIALECT

DEEL I

Promotor :
Prof. Dr. V.F. VANACKER.

Proefschrift voorgelegd aan de
Faculteit Letteren en Wijsbegeerte
voor het verkrijgen van de graad
Licentiaat in de Germaanse
Filologie door
Piet STANDAERT.

NO SYSTEMATIC ANALYSIS OF WHOLE COLLECTION

UNIVERSITEIT GENT

2014: DIGITISATION + WEBSITE

www.dialectloket.be

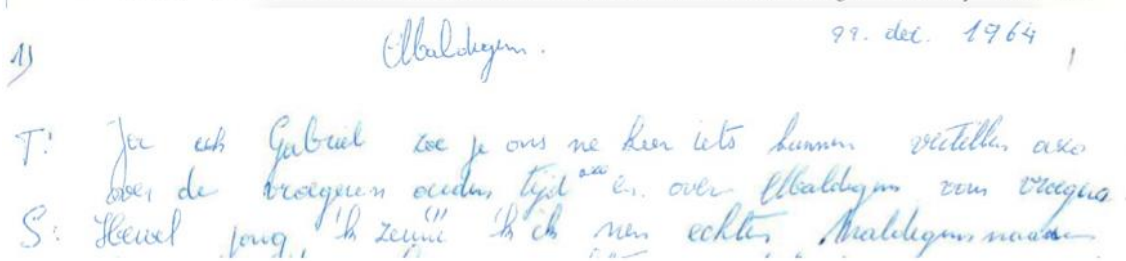
BUT

- **Of limited use for linguistic research**
  - Transcription only for 40% of collection
  - Transcription of varying quality, no unified protocol
  - No syntactic annotation

van kamere. Waarda'k goeng,ze lag in een andere kamere,maar 't was in den doo' kamere (doodkamer). Ze lag ip sterven. 'k Kom erbi,'k zie't,ze vraag no mi,'eur kleed voor an te doen en een laken voor in te draaie. 'k Zegge morgenuchtend is ze dood,'k goenk me' lakens,'t was ollemolle (allemaal) gedaan. 'k Zegge morgenuchtend is ze dood. Oo'k do 's nuchtens (Als ik daar 's ochtends) bi komme,ze leef' nog/ ja/ ze stonden erbi voor of te

L. D'r was zeker veel armoe dan in dien tijd (h)ier ?

S. Hojojo, d'r was (h)ier gee(n) werk (h)aast, newar, 't waren gelukkigen die bij nen boer mochten gaa(n) werken ... voor nen boter(h)am (h)é, maar voor gee(n) geld, zelle. Da(t) wa(s) ne gelukkige mens die da(t) mocht doen. De mensen kwamen uit Vrankrijk newar, dan ...



UNIVERSITEIT GENT

GCND-PROJECT

UNIVERSITEIT
GENT

FACULTEIT LETTEREN
EN WIJSBEGEERTE

GOALS

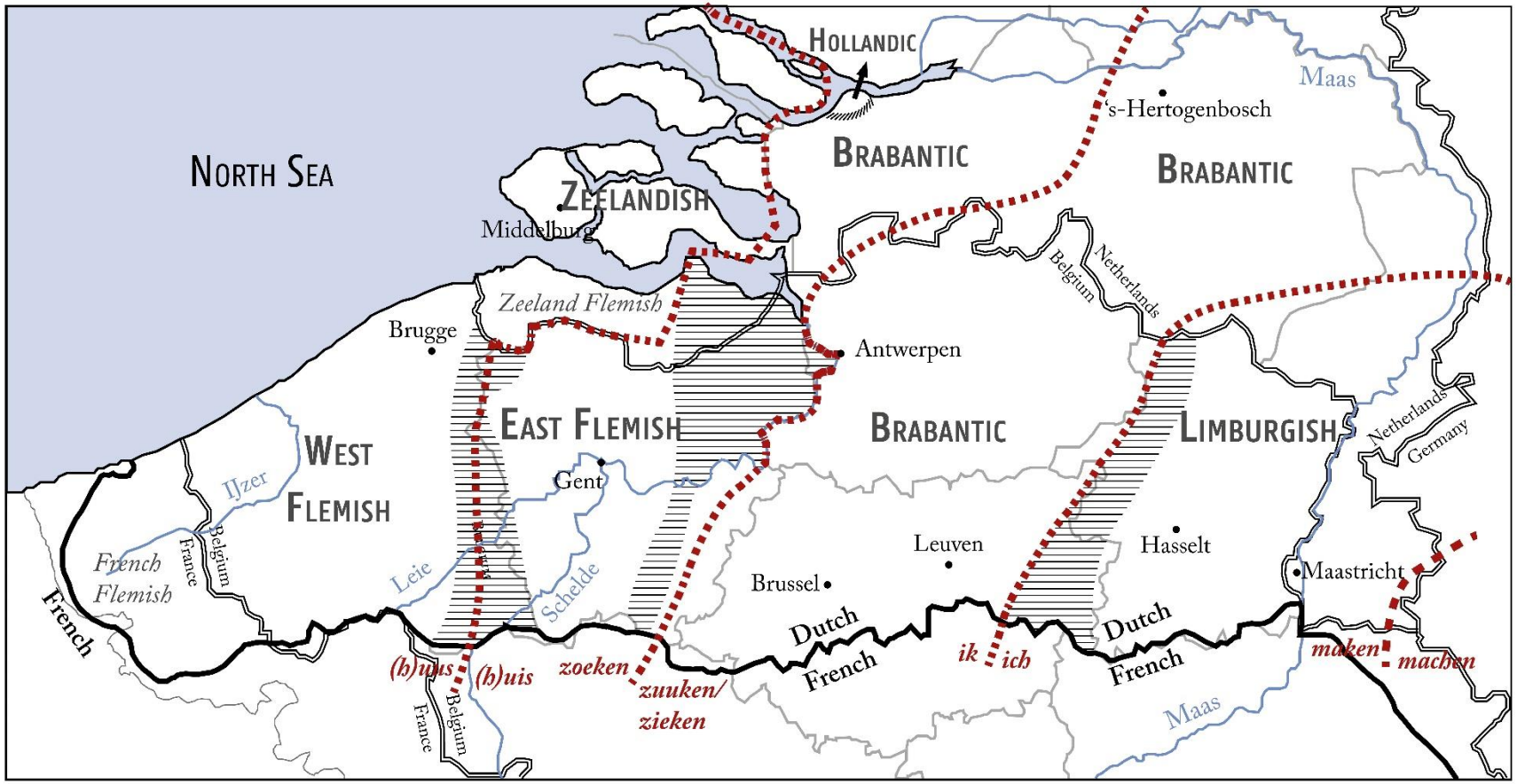
UNIVERSITEIT
GENT

GOALS

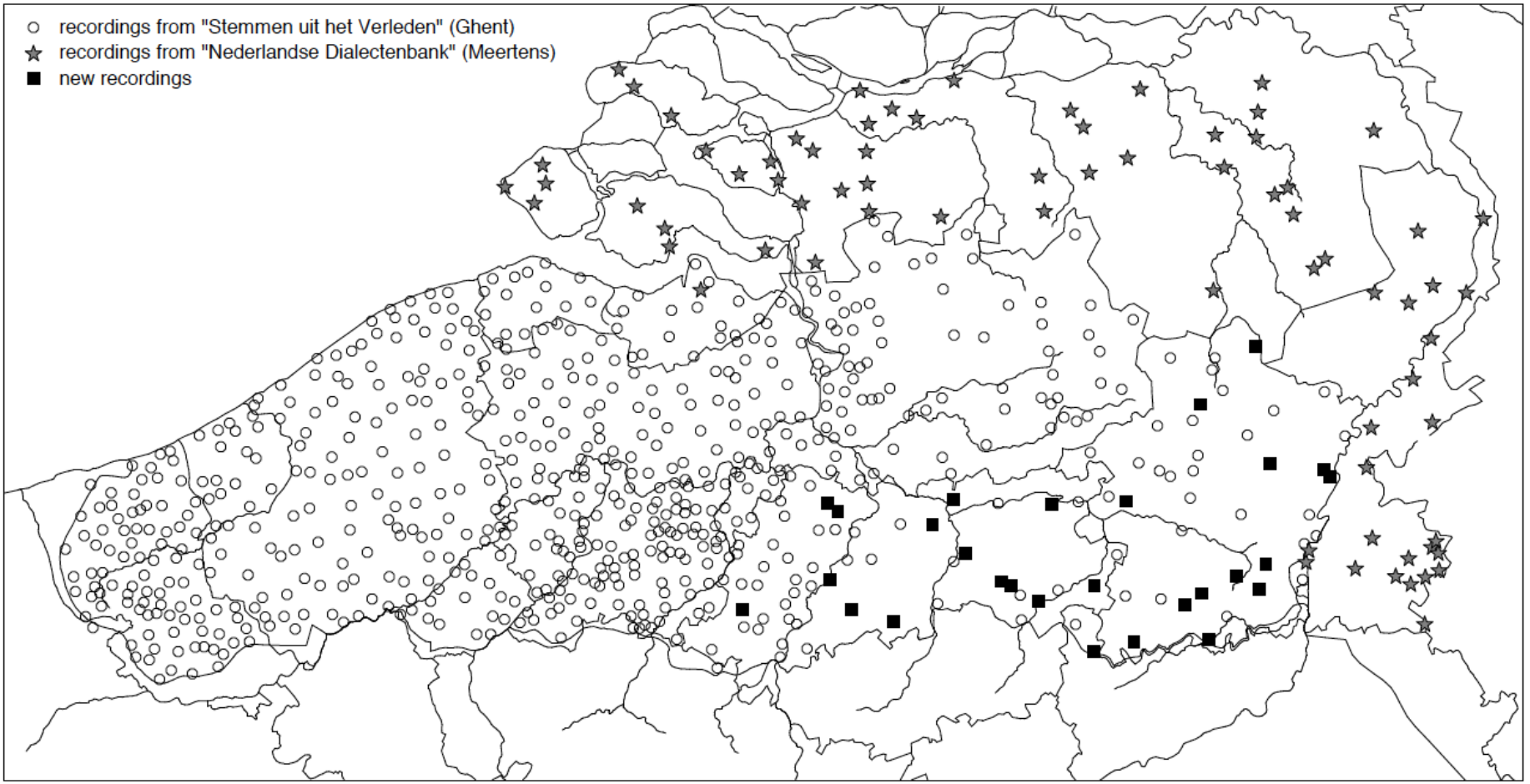
- Uniform, scientific **transcription**
- **Aligned to audio**
- **POS-tagged and parsed**
- **Searchable via online platform**

- **Enlarge collection:** + 28 new recordings > enough data on eastern dialects
- **75 recordings from the *Nederlandse Dialectenbank*** (Meertens Instituut): in order to cover a dialectologically 'logical' area

UNIVERSITEIT GENT



UNIVERSITEIT GENT



'Southern Dutch dialects'

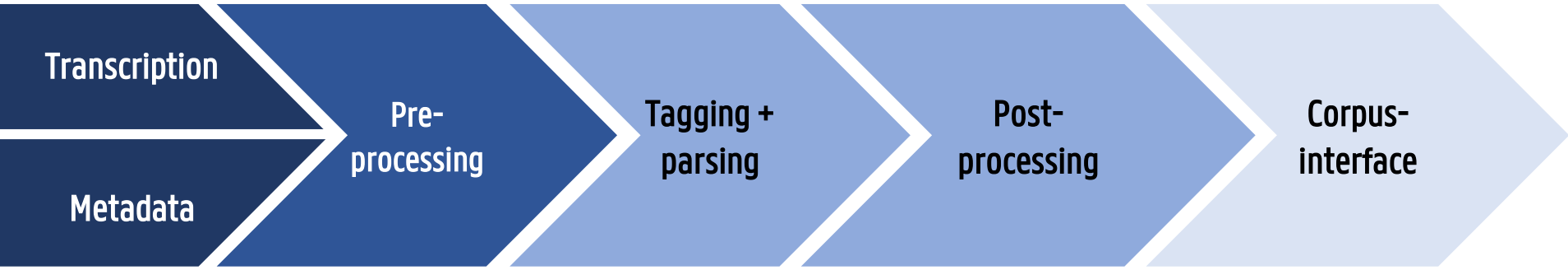
UNIVERSITEIT GENT

FACULTEIT LETTEREN
EN WIJSBEGEERTE

WORKFLOW

UNIVERSITEIT
GENT

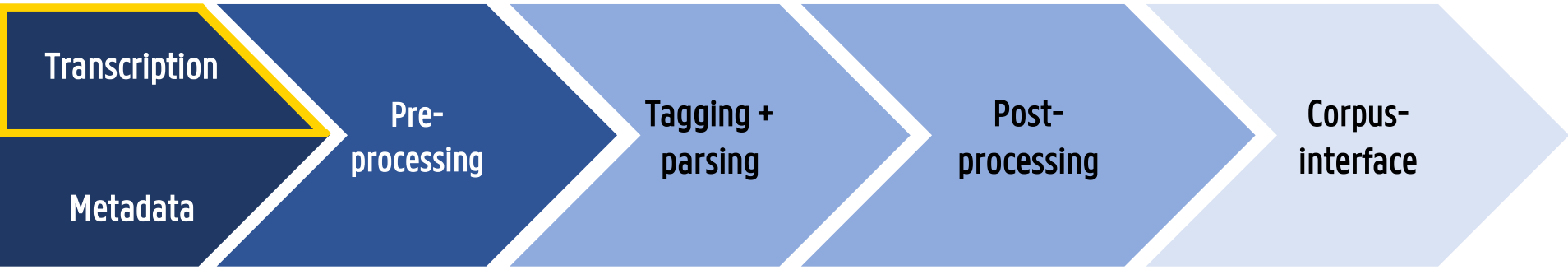
WORKFLOW



+ New recordings

UNIVERSITEIT GENT

WORKFLOW



+ New recordings

UNIVERSITEIT GENT

TRANSCRIPTION

- Difficult, labor-intensive hurdle in spoken dialect corpus building
- Tests with ASR, respeaking and forced alignment (cf. Ghyselen et al. 2020)

TRANSCRIPTION

- Decision 2019: manual transcription most time efficient
- **Forced Alignment might be useful in future for word-level alignment of audio and transcription**

UNIVERSITEIT GENT

TRANSCRIPTION

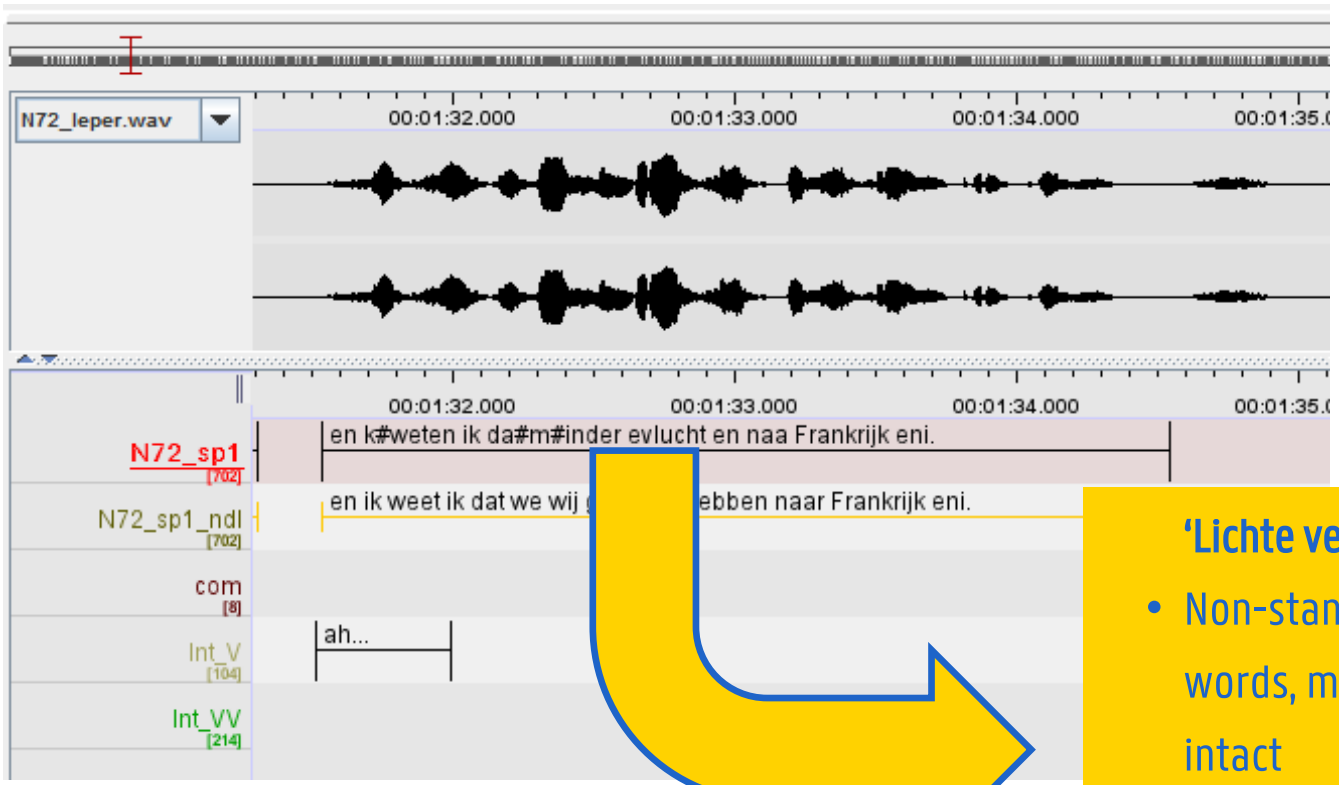
- **Manual transcription** via software package **ELAN** (free + open-source)
- **Alignment with audio:** manual segmentation on sentence level, later: possibility of forced alignment on word level
- **Protocol (Ghyselen et al. 2020): balancing act**:
  - Visualize dialect as reliably as possible

  Vs.

  - Feasability, searchability and homogeneity

UNIVERSITEIT GENT

TWO LAYERS OF TRANSCRIPTION

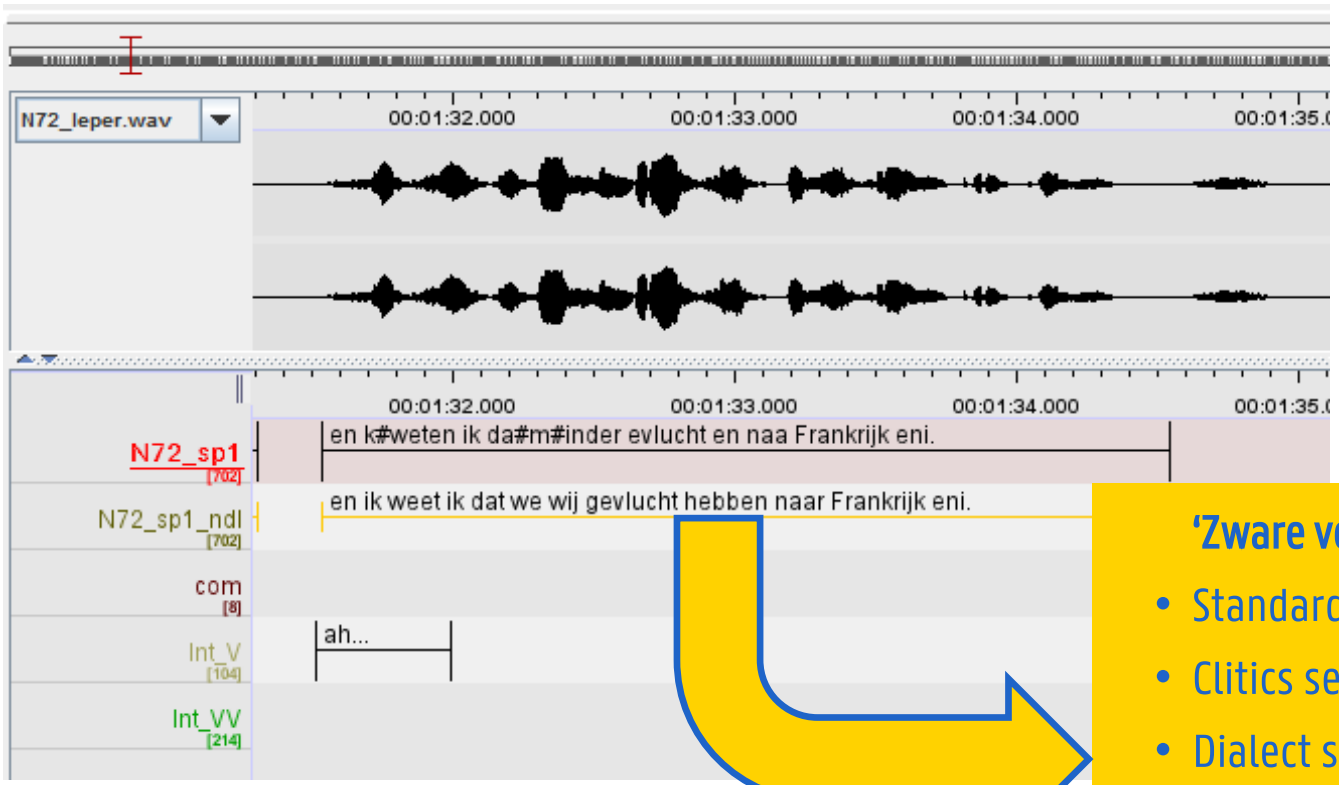


'Lichte vernederlandsing'

- Non-standard vocabulary, function words, morphology and syntax intact
- Standardization of dialectal vowels and consonants

UNIVERSITEIT GENT

TWO LAYERS OF TRANSCRIPTION



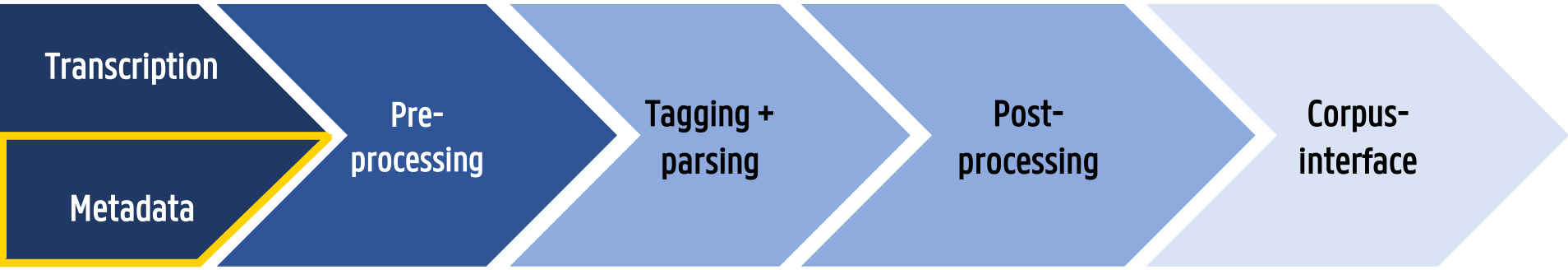
'Zware vernederlandsing'

- Standard Dutch morphology
- Clitics separated
- Dialect syntax and vocabulary preserved

UNIVERSITEIT GENT

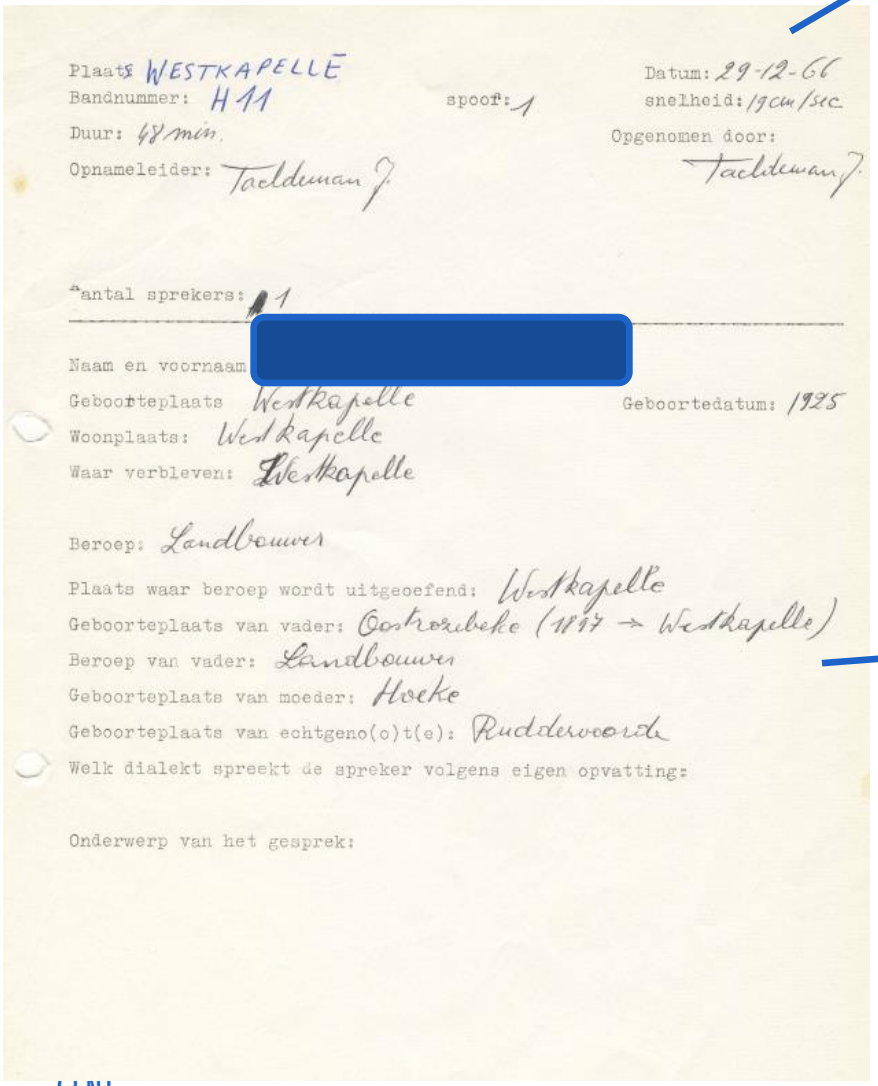

PROTOCOL > STUDENTS (AND SOME VOLUNTEERS) >
TRANSCRIPTIONS > > QUALITY CHECK BY VOLUNTEERS
(NATIVE SPEAKERS OF DIALECT)

WORKFLOW



UNIVERSITEIT GENT

METADATABASE



Plaats: WESTKAPELLE
Bandnummer: H 11
spoel: 1
Datum: 29-12-66
snelheid: 19cm/sec
Duur: 48 min.
Opgenomen door: Taeldeman J.
Opnameleider: Taeldeman J.

Aantal sprekers: 1

Naam en voornaam: [redacted]
Geboorteplaats: Westkapelle
Geboortedatum: 1925
Woonplaats: Westkapelle
Waar verbleven: Westkapelle

Beroep: Landbouwer
Plaats waar beroep wordt uitgeoefend: Westkapelle
Geboorteplaats van vader: Oostrozebeke (1947 → Westkapelle)
Beroep van vader: Landbouwer
Geboorteplaats van moeder: Hoeke
Geboorteplaats van echtgeno(o)t(e): Ruddervoorde
Welk dialekt spreekt de spreker volgens eigen opvatting:

Onderwerp van het gesprek:

Information about recording

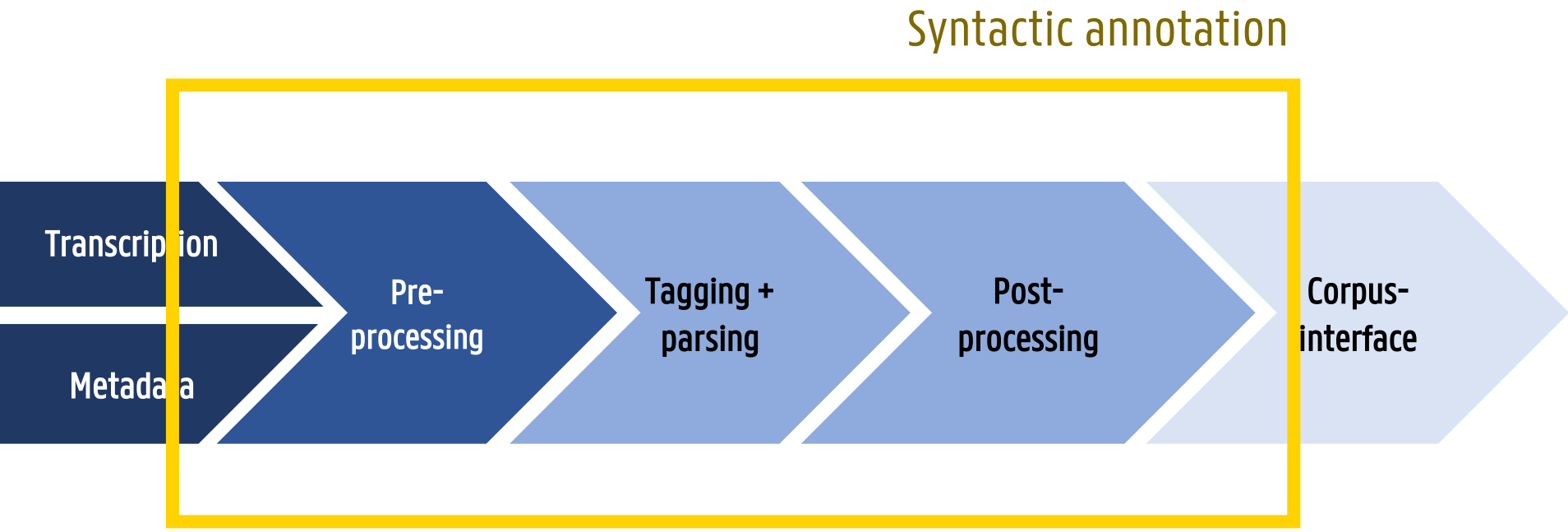
- Date
- Duration
- Interviewer
- Speakers
- …

Information about speaker

- Occupation
- Occupation parents
- Village of birth
- …

Metadatafile per speaker + per recording (linked) > standardized open responses

WORKFLOW



UNIVERSITEIT GENT

TAGGING EN PARSING

- Tests: Farasyn et al. (2022)
- Using **Alpino parser**: **dependency parser** for Dutch (Van den Bosch et al. 2007)
- Mainly trained on **written (Standard) Dutch**
- Problem: difficulties with **constructions typical for dialect**
  - *ik **en** heb dat niet geweten*
    'I NOT have that NOT known'
  - ***Ik** heb **ek ik** dat niet gezegd.*
    'I have I I that not said'.
- **Some manual help is necessary**

UNIVERSITEIT GENT

https://www.let.rug.nl/vannoord/alp/Alpino/

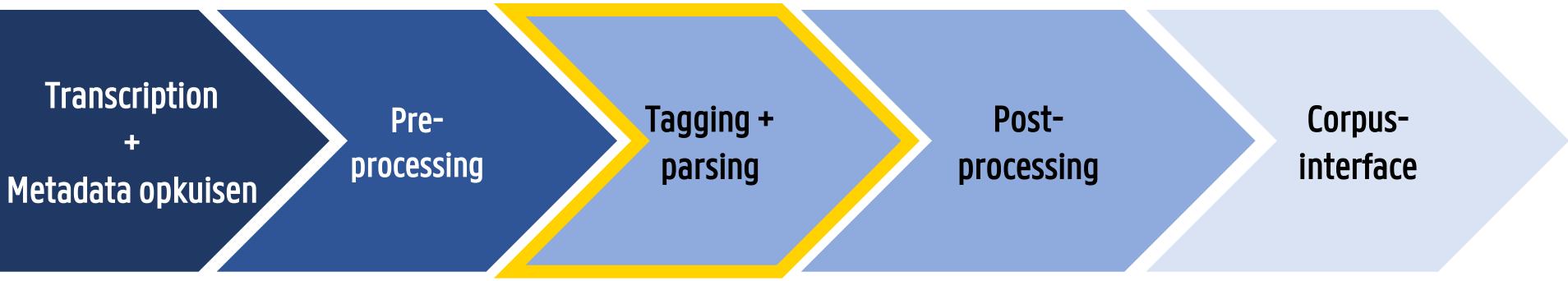
PREPROCESSING

Add commands to

- skip certain items **[@skip]**
- treat one item as if it were another**[@alt]**
- Add 'phantom words' that make the analysis easier [@phantom]
- **…**

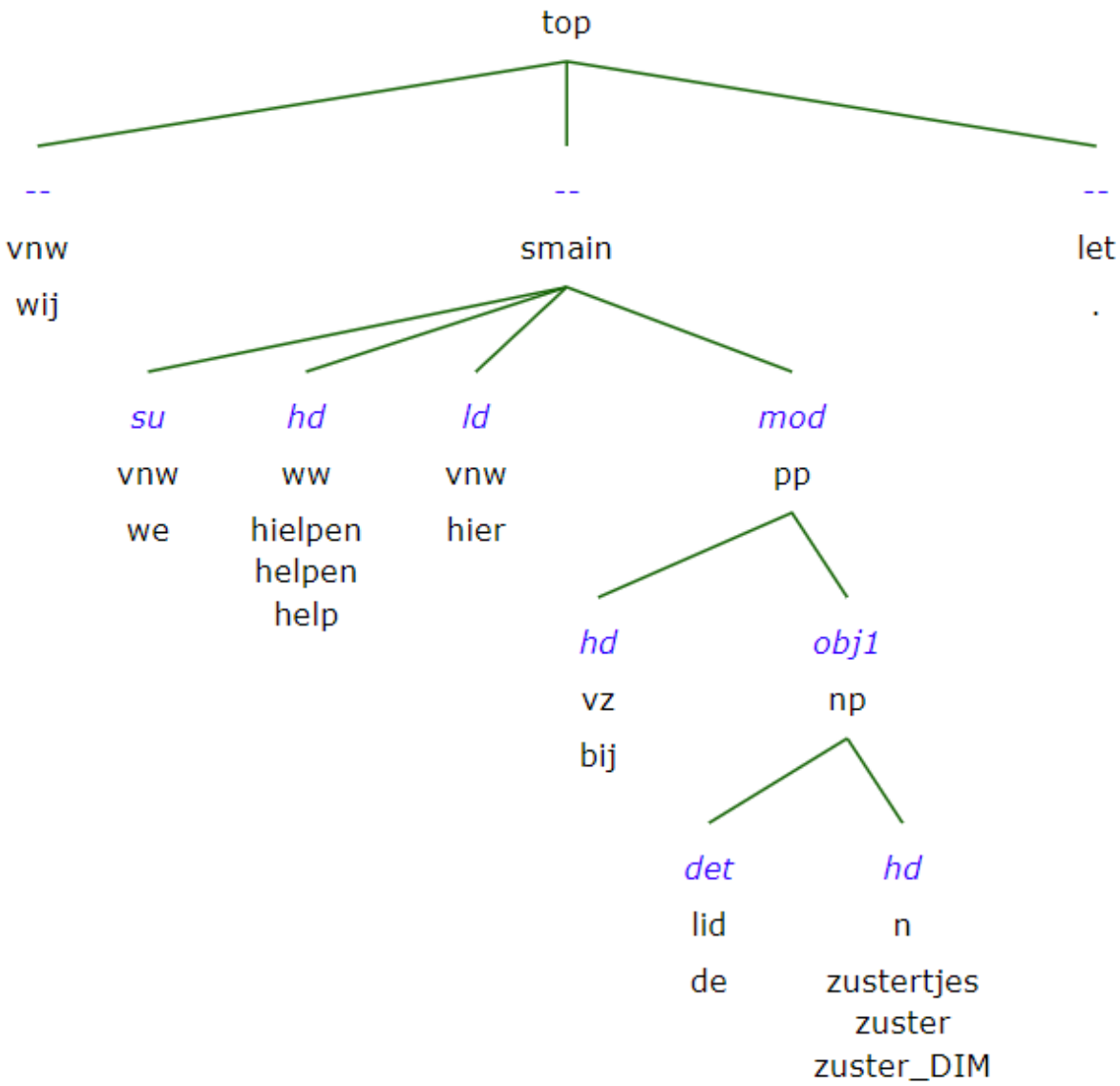
H117p_1--H117_1_1--0062|we hielpen [@skip wij] hier bij de zustertjes .
H117p_1--H117_1_1--0063|en [@skip uh] ik [@skip en] stond niet mager maar ik stond toen nog vetter .

WORKFLOW



UNIVERSITEIT
GENT

we hielpen wij hier bij de zustertjes .

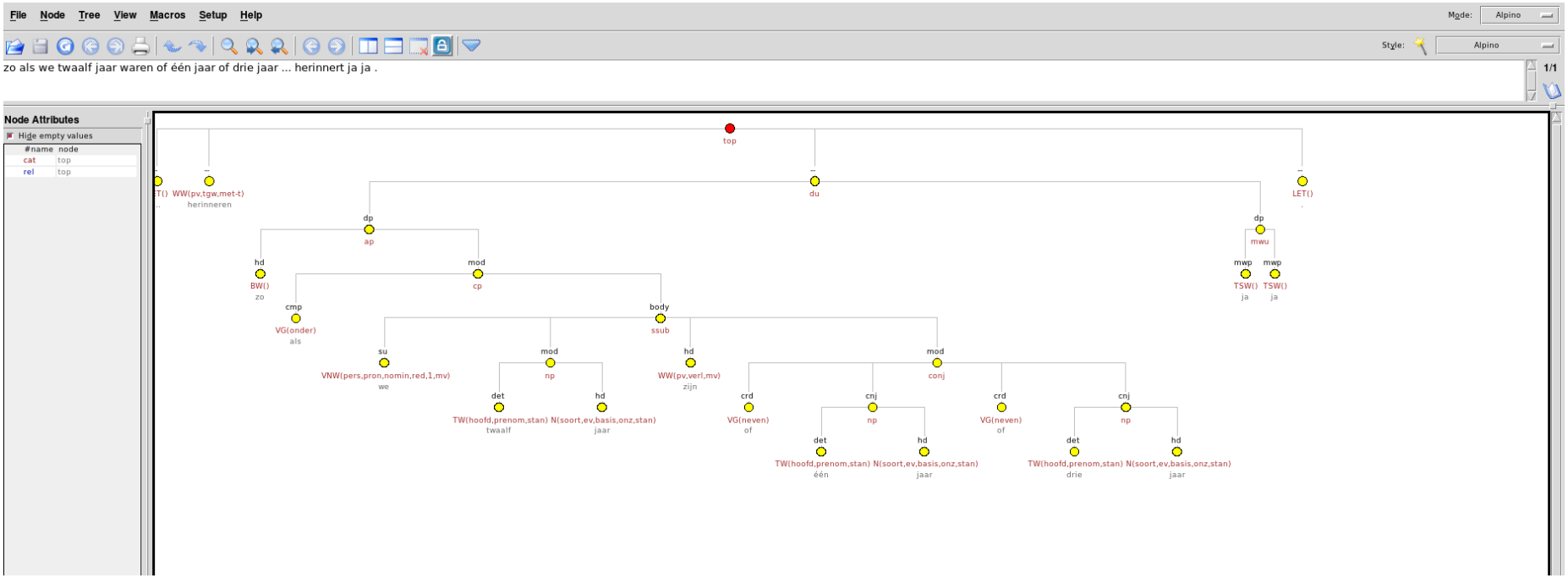


UNIVERSITEIT GENT

WORKFLOW

UNIVERSITEIT
GENT

TREE EDITOR TRED 2.0



UNIVERSITEIT GENT

https://ufal.mff.cuni.cz/tred/

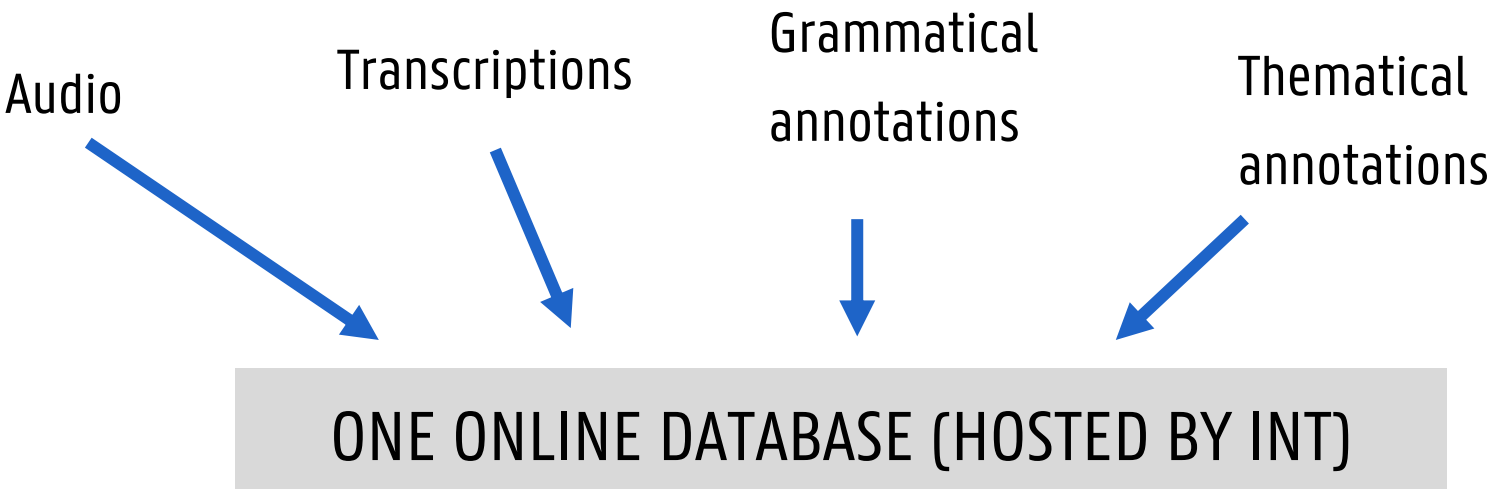
POSTPROCESSING

- Elaborated manual
- Only advanced students
- ‘Problem database’

UNIVERSITEIT
GENT

WORKFLOW

UNIVERSITEIT
GENT



UNIVERSITEIT GENT

/instituut voor de Nederlandse taal/

INTERFACE

- Cf. **Corpus Hedendaags Nederlands** https://chn.ivdnt.org/
- **In line with CLARIN-philosophy** ('Common Language Resources and Technology Infrastructure')
- **Separate search modules**: *simple – extended – advanced – exp*
- **Corpus query language (CQL)**
- Searchable on text/structure/content
- Filters for place/region/age/content/…
- Results: text + audio + tree

UNIVERSITEIT GENT

FACULTEIT LETTEREN
EN WIJSBEGEERTE

LOOKING FORWARD

UNIVERSITEIT
GENT

GCND

- Metadatabase almost ready
- INT working on interface
- 662 (91%) transcribed
- 270 (42%) transcriptions doublechecked
- 63 (8,6%) preprocessed
- 10 (1,3%) postprocessed

Parsing and tagging more labour-intensive than estimated in advance

Priorities

(1) Transcription

(2) Preprocessing and postprocessing small regionally balanced subset of recordings

UNIVERSITEIT GENT

FOLLOW-UP PROJECT (PROPOSAL IN PREPARATION)

- **Regional expansion of corpus** (in collaboration with Meertens)
- **Parse more data**
- **Speed up the process by**:
  - Experimenting with ASR
  - Retraining parser

Proposal in preparation

Partners:
- Ghent University:
- Meertens Institute
- KU Leuven
- INT

UNIVERSITEIT GENT

GCND:

Unparalleled corpus of spoken Dutch dialect data

> Invaluable opportunities for research on dialect syntax

UNIVERSITEIT GENT

FACULTEIT LETTEREN EN WIJSBEGEERTE

ANNE-SOPHIE GHYSELEN
annesophie.ghyselen@ugent.be

MELISSA FARASYN
melissa.farasyn@ugent.be

ANNE BREITBARTH
anne.breitbarth@ugent.be

Thanks for your attention

Questions? Remarks?

UNIVERSITEIT GENT

gcnd@ugent.be

REFERENCES

Farasyn, M., A.S. Ghyselen, J. Van Keymeulen and A. Breitbarth. 2022. "Challenges in tagging and parsing spoken dialects of Dutch." Journal of Historical Syntax 6: 4-11.

Ghyselen, A.-S., A. Breitbarth, M. Farasyn, J. Van Keymeulen & A. Van Hessen. 2020. "Clearing the transcription hurdle in dialect corpus building : the corpus of Southern Dutch dialects as case-study." Frontiers in Artificial Intelligence 3: 1-17.

Ghyselen, A.-S., J. Van Keymeulen, A. Breitbarth & M. Farasyn. 2020. "Het transcriptieprotocol van het Gesproken Corpus van de Nederlandse Dialecten (GCND) " Handelingen van de koninklijke commissie voor toponymie & dialectologie 92: 83-115.

Van den Bosch, A., B. Busser, S. Canisius & W. Daelemans. 2007. "An efficient memory-based morphosyntactic tagger and parser for Dutch." LOT Occasional Series 7. 191–206.