

STEMMEN UIT HET VERLEDEN ONTLEDEN: HET GESPROKEN CORPUS VAN DE (ZUIDELIJK-)NEDERLANDSE DIALECTEN (GCND)

Het project:

- Geparst (=syntactisch geannoteerd) corpus van spontaan gesproken (historische) zuidelijk-Nederlandse dialecten – een groep dialecten die een aantal typologische bijzonderheden vertonen, die enerzijds een ouder taalstadium bewaren en anderzijds een aantal innovaties hebben doorgevoerd, en in beide gevallen verschillen van de standaardtaal.
- Steunt op een lange traditie in de dialectologie in Gent

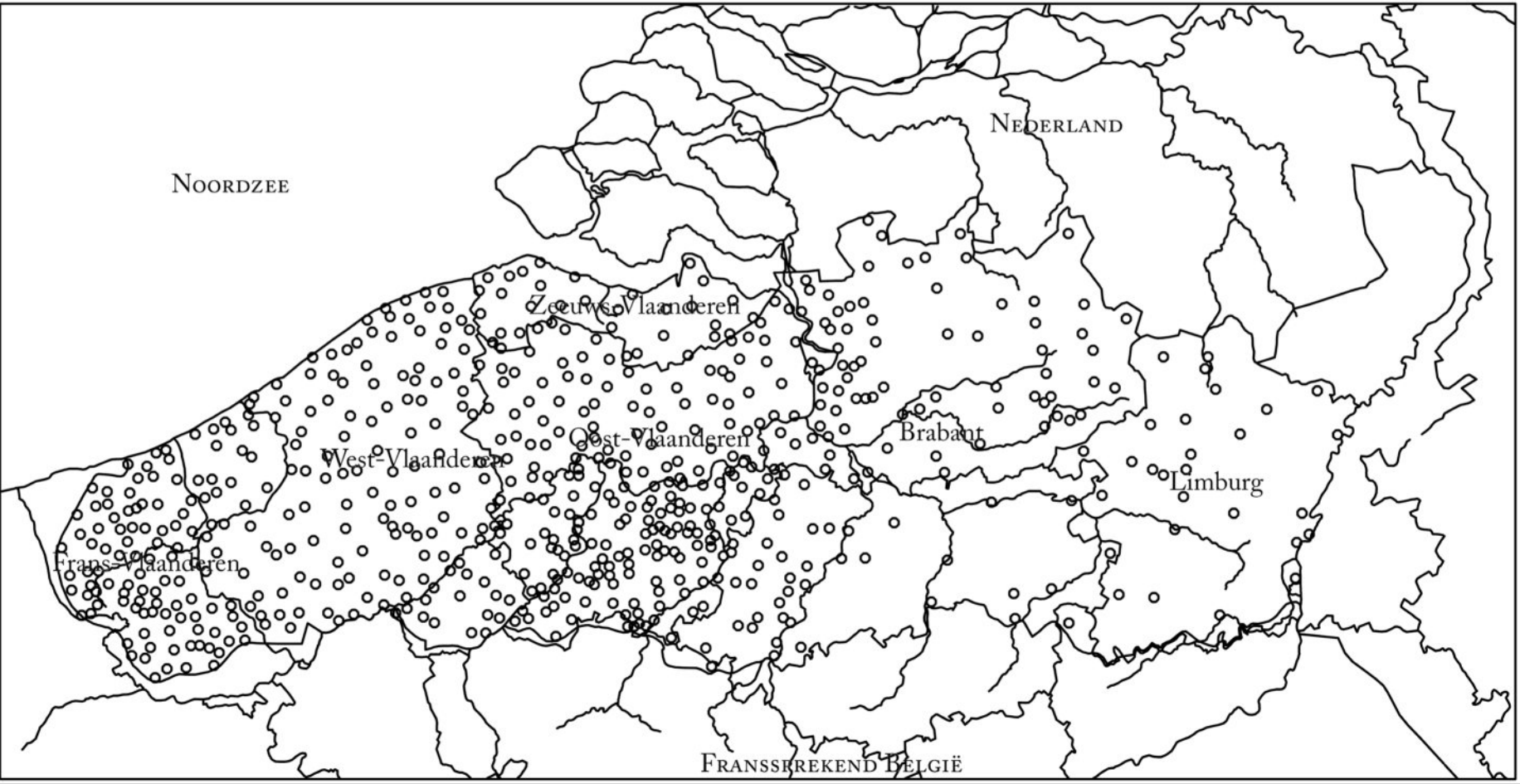
I. De voorgeschiedenis

Dialectologie in Gent

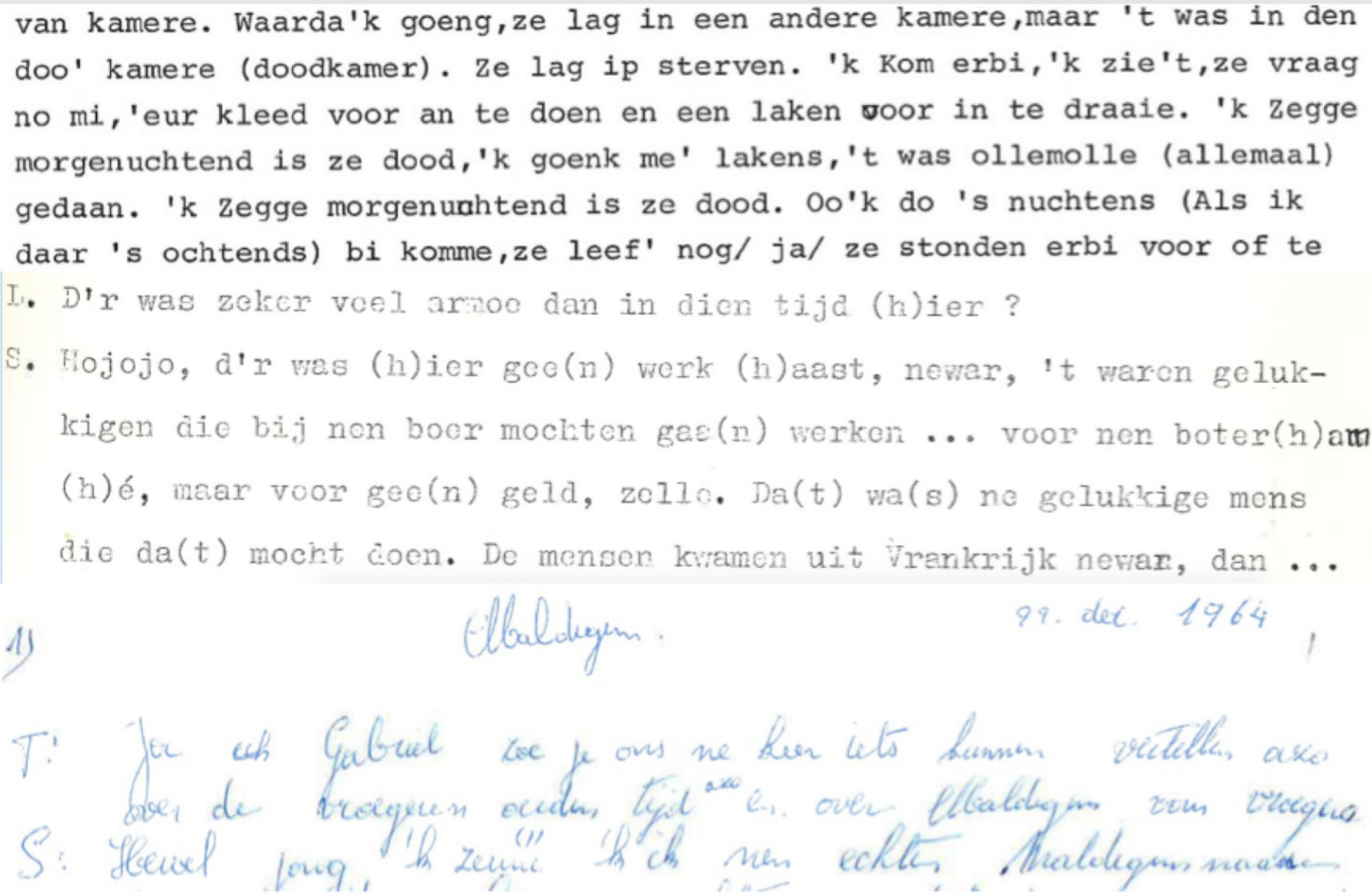
- 1923-: verzameling van de fonetische transcripties van 141 zinnen in 1956 plaats in de Nederlandse en Friese taalgebieden door E. Blancquaert – basis voor de *Reeks Nederlandse Dialectatlassen*, nu online beschikbaar: <https://www.dialectzinnen.ugent.be>
- 1972-2018: *Woordenboek van de Vlaamse Dialecten* (W. Pée), nu online beschikbaar: <http://www.e-wvd.be>
- Gentse medewerking aan
 - de Fonologische Atlas van de Nederlandse Dialecten (FAND) (J. Taeldeman) <http://www.meertens.knaw.nl/mand/database/>
 - de Syntactische Atlas van de Nederlandse Dialecten (SAND) (M. Devos, G. De Vogelaer) <http://www.meertens.knaw.nl/sand/>
- 2017-: FWO-Infrastructuurproject *Dictionary of the Southern Dutch Dialects (DSDD). An integrated lexicological database for the southern Dutch dialects* (J. Van Keymeulen; Van Keymeulen et al. 2018; ter perse)

Stemmen uit het verleden

- collectie van 783 bandopnamen uit 550 plaatsen in Frans-, West-, Zeeuws- en Oost-Vlaanderen, Vlaams-Brabant en Limburg, gemaakt tussen 1964 en 1975 (V.F. Vanacker), online te beluisteren op www.dialectloket.be



- van taalkundig én ethnologisch belang: ongeziene registratie van traditionele dialecten + unieke inzicht in maatschappij van 1e helft 20e eeuw
- transcripties van 318 van die opnames van zeer uiteenlopende kwaliteit, geen geünificeerd transcriptieprotocol:

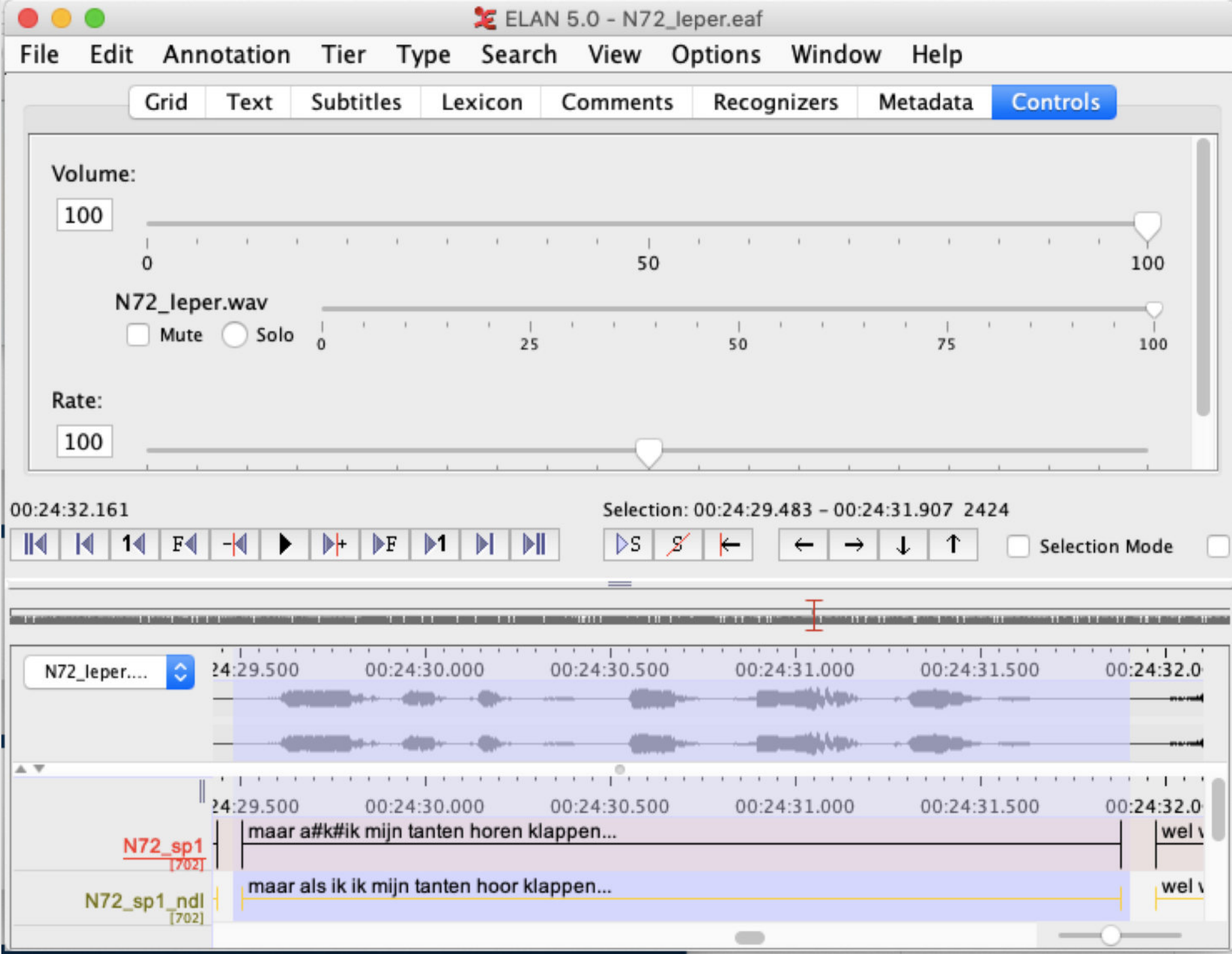


Waarom:

- Wegens het ver gevorderde **dialectverlies** in Vlaanderen **dringt de tijd** om de unieke collectie geluidsopnames te ontsluiten voor de volgende generaties.
- Het GCND is gebaseerd op **spontaan gesproken taal**, **geëliciteerde data** (vragenlijsten). Daardoor is er **geen preselectie** van te onderzoeken fenomenen. Dat maakt nieuwe, ook toevallige, ontdekkingen mogelijk.

II. Pilootfase (2018-2019)

- Transcriptieprotocol** (Ghyselen et al. in voorbereiding)
 - transcriptie in ELAN in twee lagen: één dicht bij dialect, één dicht bij de standaard
- Laag 1: maar a#k#ik mijn tanten horen klappen
Laag 2: maar als ik ik mijn tante hoor klappen



- NLP-experimenten** (Breitbarth et al. geaccepteerd-a,b)
 - probleem: bestaande POS-taggers zijn vaak getraind op (geschreven) Algemeen Nederlands
 - sommige zijn getraind op het Corpus Gesproken Nederlands (CGN), nog steeds dicht bij de standaardtaal
 - gevolg: lage accuratesse
 - tests: drie NLP-tools, 10 plaatsen (FV, WV, ZV, OV, VB, BL, en overgangsgebieden), accuratesse in %

plaats	TreeTagger	LeTS	FROG
Oudenburg (H24)	94.2	92.9	95.2
Maldegem (I154)	94.7	94.4	97.3
Westdorpe (I166)	90.9	93.0	98.8
Sint-Niklaas (I175)	96.3	91.4	94.6
Gent (I241)	91.8	88.4	92.9
Ieper (N72)	90.4	82.5	92.6
Hardifort (N94)	90.1	88.5	93.8
Sint-Joris-Weert (p130)	91.8	90.5	89.3
Uikhoven (Q013)	95.9	92.8	95.9
Totaal	92.9	90.5	94.5

- Eerste exploraties (voorlopige dataset)**
 - bvb. (oorspronkelijk) ontkennend *en*. *en* als discoursmarkeerder (Breitbarth & Haegeman 2015) naast "Middel Nederlands" gebruik (Breitbarth et al. geaccepteerd-a):

- met zijn beste kleren aan ... je had dien een keer moeten **en** zien (N42 Pittem)
- ik **en** weet of dat nu nog veel meer gedaan werd (O265 Ronse)

- Inhoudelijke annotatie**
 - eerste stappen in de richting van een systeem voor het taggen van inhoudscategorieën, bijv. op basis van het bestaande Library of Congress Subject Headings (LCSH)-systeem of de Art & Architecture Thesaurus (AAT)

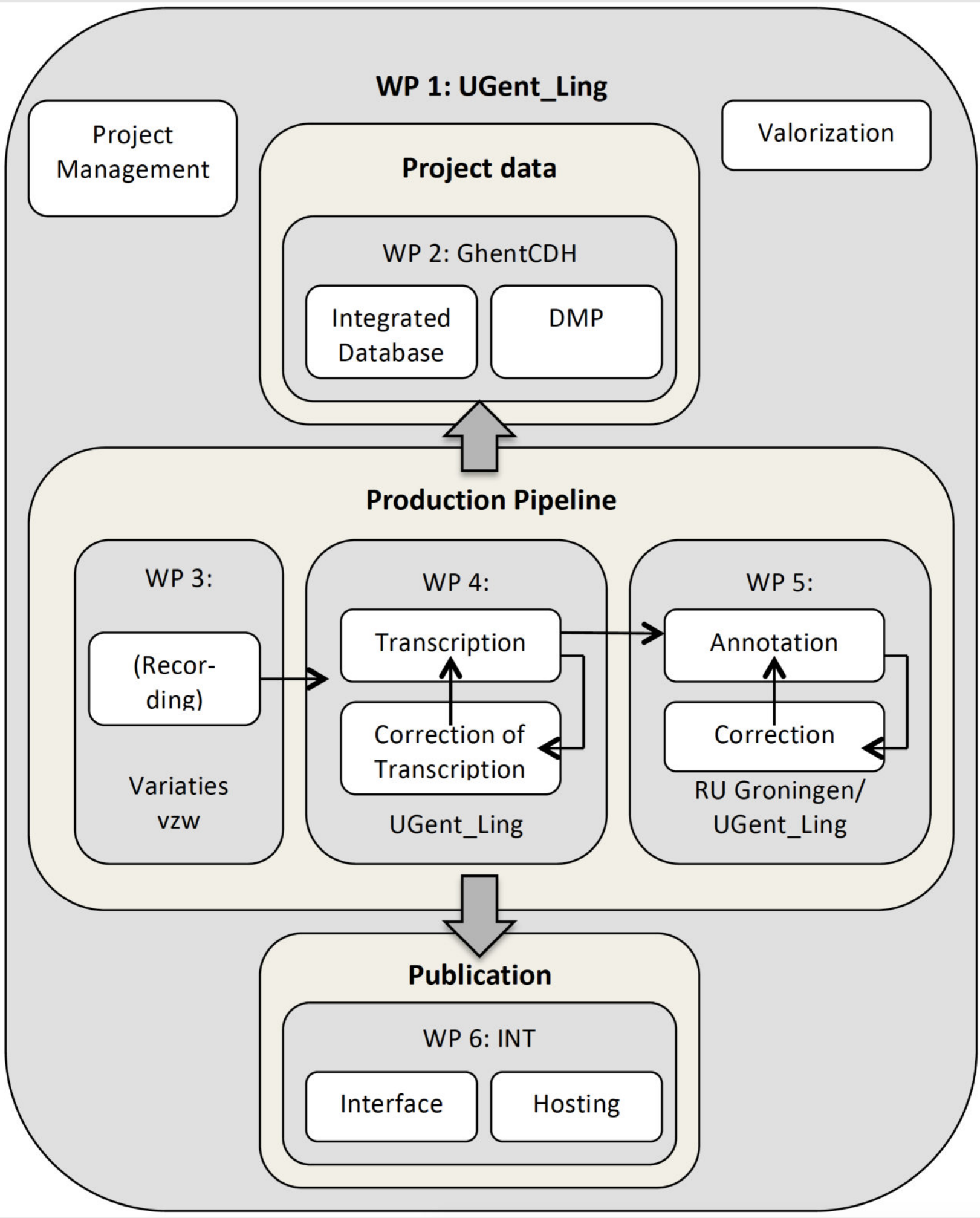
Wat:

- Transcriptie van opnames van spontaan gesproken dialect
- Annotatie van woordsoort & morfologie (POS-tagging)
- Parsing (syntactische annotatie)

III. De toekomst

FWO-infrastructuurproject (in voorbereiding):

- Internationale samenwerking met:
 - UGent Taalkunde: onderzoeksgroepen Dialing & GLIMS
 - GhentCDH
 - Variaties vzw.
 - Instituut voor de Nederlandse Taal (INT)
 - Meertens Instituut
 - Rijksuniversiteit Groningen



Het plan:

- Transcriptie en annotatie van opnames uit een fijnmazig plaatsennet uit het hele zuidelijk-Nederlandse dialectgebied
- Duurzame, uitbreidbare infrastructuur @INT
- Doorzoekbaar voor **syntactische structuren** (boompjes), met een op **GrETEL** voortbouwende interface, alsook **metadata**, met een op **OpenSoNaR+** voortbouwende interface

(Inter)nationale samenwerkingen (in voorbereiding):

- Uitbouw van de zoekmogelijkheden door toevoeging van karteringsfunctionaliteit i.s.m. de DSA Marburg (REDE SprachGIS)
- WOG met o.m. KU Leuven, U Antwerpen, het Meertens Instituut, Cambridge, Padua, Glasgow, Tromsø, Yale, CUNY
- Het FWO-infrastructuurproject "A Respeaking and Collaborative Game-Based Approach to Building a Parsed Corpus of European Spanish Dialects", gebaseerd op delen van het COSER

Breitbarth, A. & L. Haegeman. 2015. 'En' en is niet wat we dachten: a Flemish discourse particle. *MIT Working Papers in Linguistics* 75: 85–102.

Breitbarth, A./M. Farasyn/A.-S. Ghyselen/L. Haegeman/J. Van Keymeulen. Geaccepteerd(a). *Ge had dien e keer moeten en zien!* Neue Erkenntnisse zum Gebrauch der Partikel *en* im Gesproken Corpus van de (Zuidelijk-)Nederlandse Dialecten. In Speyer, Augustin & Anne-Kathrin Balo (Hrsg.): *Syntax aus Saarbrücker Sicht 4 – Beiträge der SaRDiS-Tagung zur Dialektsyntax. Zeitschrift für Dialektologie und Linguistik – Beihefte (ZDL-B)*. Stuttgart: Steiner-Verlag.

Breitbarth, A./M. Farasyn/A.-S. Ghyselen/J. Van Keymeulen. Geaccepteerd(b). Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten. *Handelingen van de Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis (KZM)*.

Ghyselen, A.-S./A. Breitbarth/M. Farasyn. In voorbereiding. Clearing the transcription hurdle in dialect corpus building: The corpus of Southern Dutch dialects as case-study. Ms. Ghent University.

Van Keymeulen, J., V. De Tier, A. Breitbarth, A. Ghyselen & M. Farasyn. Ter perse. Het Corpus "Stemmen uit het Verleden" van de Universiteit Gent. *Volkskunde*.

Van Keymeulen, J./S. Chambers/V. De Tier/J. de Does/K. Depuydt/T. Schoonheim/R. Vandenbergh/L. Hallebaut. 2018. Sustaining the Southern Dutch Dialects: the Dictionary of the Southern Dutch Dialects (DSDD) as a case study for CLARIN and DARIAH. In: Skadin, I. & M. Eskevich (eds.), *Proceedings of the CLARIN Annual Conference 2018*, 132–136.

Van Keymeulen, J./V. De Tier/R. Vandenbergh/S. Chambers. Ter perse. The Dictionary of the Southern Dutch Dialects. Designing a virtual research environment for digital lexicological research. *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*.