FACULTY OF ARTS
AND PHILOSOPHY

Department of Linguistics
Research group DiaLing

INTO THE WILD

Interpretation of Ecologically Valid Data

Claudia Crocco, Anne Breitbarth, Alexandra Simonenko,
Kim Groothuis, Giuseppe Magistro & Giovanni Leo

4th LW Research Day *From Source to Understanding*, 27 November 2024

GHENT
UNIVERSITY

INTRODUCTION

"There is no doubt that studying individual organisms in isolation has helped us to understand the basic features of how individual species make their living. But this approach also biased our understanding of how microbial communities function. It is akin to extrapolating the behavior of an African cichlid in my aquarium to their behavior in the lakes in which they live. An aquarium is not a natural environment. Neither is a Petri dish nor a test tube of liquid media containing nutrients thousands of times more concentrated than in the ocean or lakes."

(Paul G. Falkowski. 2015. *Life's Engines. How Microbes Made Earth Habitable*. Princeton/Oxford: Princeton University Press.)

WRITTEN VS. SPOKEN LANGUAGE

- Linguistic analysis predominantly based on written/transcribed data
- Historical roots: Debate on the “better” form of language (*Phaedrus* by Plato)

- Advocacy for spoken language: Primary and ecologically valid manifestation of language

ECOLOGICAL VALIDITY

- Examples of controlled data:
  - Syntax: Grammaticality/acceptability judgments through questionnaires
  - Phonetics/Phonology: Laboratory-controlled, elicited, or read materials

Ecologically valid data: Data grounded in and relevant to real-world settings.

(©joart lin, Flickr)

(©bjornman, Flickr)

INTO THE WILD

- Challenges of studying dialects and under-documented languages:
  - Non-standard varieties and unnamed languages
  - Speakers with limited formal education: difficulty performing metalinguistic tasks
- Limitations of rigid methods: Overlooking spontaneous linguistic behavior



TRADE-OFFS IN DATA COLLECTION

- Rigid methods:
  - \+ Quantitatively analyzable
  - — Risk of misinterpreting as “authentic” language
- Ecological methods:
  - \+ Capture authentic behavior
  - — time-consuming; less structured data (how to interpret?)

CONSEQUENCES FOR INTERPRETATION

- Blind spot: Controlled/artificial data treated as representative of the language
- Elicitation can only find what was asked for
- Analogy: Language as a “Petri dish” phenomenon

(©lornaxu, Flickr)

DEFINING NATURAL AND ECOLOGICAL DATA

- Real-world data
- Language with a real communicative purpose
- Speech produced outside the lab for a real audience

RESEARCH FOCUS

- Exploring ecological validity in linguistic data
- Methodological goal:
  Combine ecological data with both qualitative and quantitative analyses
- Focus:
  - Non-standard/sub-standard dialectal varieties
  - Diachronic varieties

CASE STUDIES & FUTURE DIRECTIONS

1. Overview of ongoing research and infrastructure projects:
   - CAUSALITY (ERC)
   - FWO research projects: *Intonation and Rhythm of Dialects and Italian, Pragmatic Expletive Pronouns in the Dialects of Southern Italy*
   - GCND(+)
   - WuG-corpus, …
2. Insights into methodologies for enhancing ecological validity
   - Pyrlato

UNIVERSITAT DE BARCELONA

Estudios de Fonética Experimental
Journal of Experimental Phonetics
ISSN: 1575-5533 · ISSN-e: 2385-3573

EFE Estudios de Fonética Experimental

Pyrlato: A novel methodology to collect real-world acoustic data

Giuseppe Magistro [1]
Claudia Crocco [2]
[1] Ghent University (Belgium)

DOI: 10.1344/efe-2023-32-243-254
Corresponding address: giuseppe.magistro@ugent.be
Received: 12/09/2023 Accepted: 07/11/2023 Published: 28/11/2023
Magistro, G., & Crocco, C. (2023). Pyrlato: a novel methodology to collect real-world acoustic data. *Estudios de Fonética Experimental*, *32*, 243–254. https://doi.org/10.1344/efe-2023-32-243-254

THE WEB AS A LINGUISTIC RESOURCE: *PYRLATO*

- Internet as a vast corpus for written language
- Accessible and low-cost
- Source for Large Language Models (LLMs)
- BUT: web and social media are also invaluable sources of real-world speech

```
def checker():
  print("Now checking the patterns...")
  for string in pattern:
    if len(string.split()) == 1:
     for id, segments in dict_trans.items():
       list_segments = segments.get('segments')
       list_words = list(map(lambda x: x.get("words"), list_segments))
       for item in list_words:
        for iterate, sub_dictionary in enumerate(item):
         to_be_found = f' {sub_dictionary.get("word")} '
         if re.search(string, to_be_found):
           begin = lambda x: x.item, np.float64(sub_dictionary.get("start"))
           finish = lambda x: x.item, np.float64(sub_dictionary.get("end"))
           mytuple = (to_be_found, begin[1], finish[1])
           duration = mytuple[2] - mytuple[1]
           print(f'{id}: {mytuple}')
           ending = f'Pyr_{iterate}_{string}_{id}_.{format}'
           ffmpeg_extract_subclip(id, mytuple[1]-duration/3, mytuple[2]+dura
           print(mytuple[1])
```

HOW DOES *PYRLATO* WORK?

- YouTube search in the metadata to obtain relevant videos
- Audio files are extracted and transcribed using WhisperX, obtaining time stamps
- The desired string is searched through the transcriptions
- If there is a match, the audio is trimmed using the time stamps

AN EXAMPLE

- Context: In the project *CAUSALITY*, we are investigating the historical development of definite determiners in Germanic languages.
- Research topic: Many languages distinguish between two types of determiners, strong ones and weak ones. Dutch is one of these languages.

(1) (context: *er is een kat in de kamer*)
Kunt u alstublieft de kat eten geven?

(2) *Toen ik gisteren terug thuiskwam, zag ik een kat.*
De volgende dag zat die kat er nog steeds.

AN EXAMPLE

- Research question: In German, weak determiners are morpho-phonologically reduced (e.g. *vom*) in comparison to strong ones (*von dem*) —do we find a similar contrast in Dutch?
- Problem 1: While German reduction is more transparent (the ortographic transcription points explicitly to reduction), for Dutch we need to investigate reduction in spoken data.
- Problem 2: If we ask speakers to read these words or speak in front of microphones, they produce careful and hyperarticulated speech.
  Our data might not be ecologically valid and we might not observe reduction.
- Solution: *Pyrlato*

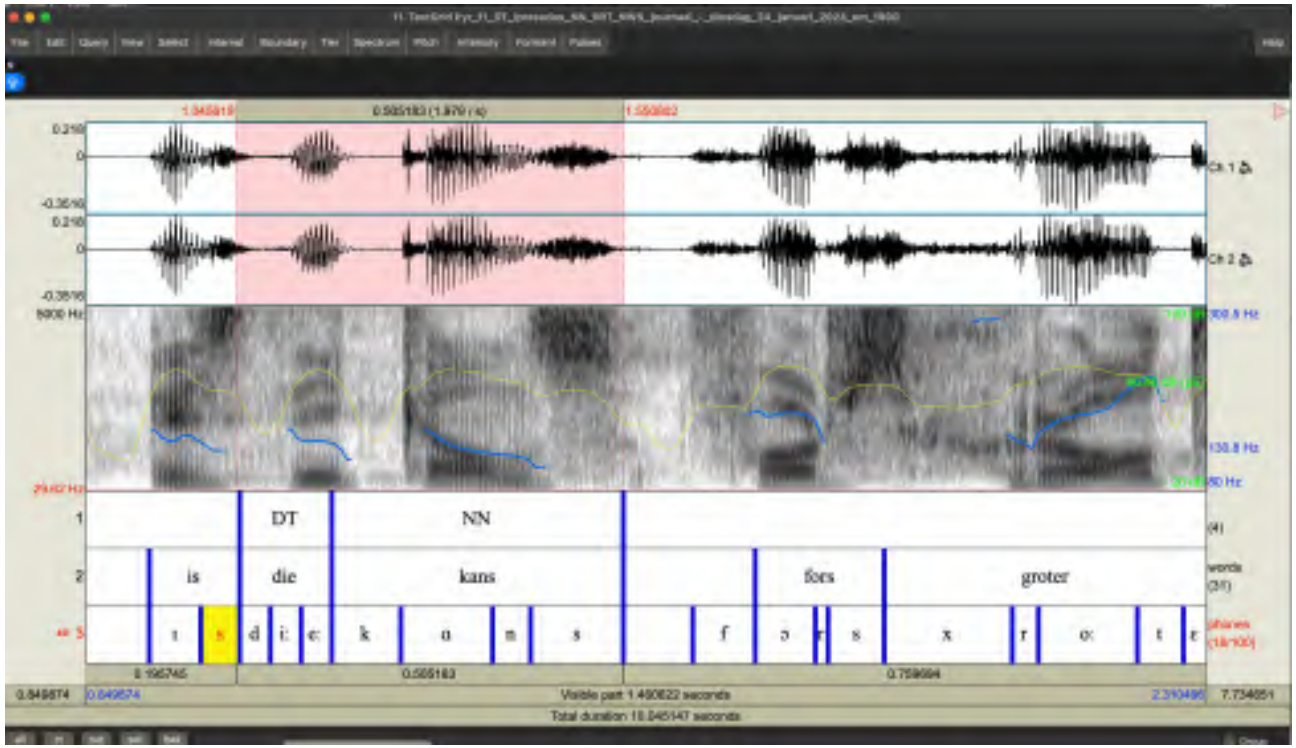
LET'S SEE *PYRLATO* IN ACTION!

- Input questions:

```
What source do you want to use? youtube
Which search keyword do you want to use? Vlaanderen
Do you want to run a demo? n
Which language are you interested in? Dutch
Which search mode do you want to use? syntax
Write here your query: DT iprecedes NN
Do you need speaker's diarization? y/n n
Do you need Praat textgrids? y/n n
Now Searching YouTube videos
Vlaanderen scoort slecht bij Europese toeristen
Tom Waes volgt een cursus West-Vlaams | Reizen Waes: Vlaanderen
Wim - 'Jealousy' | Blind Auditions | The Voice Van Vlaanderen |
Jade - 'Homesick' | Blind Auditions | The Voice Kids | VTM
```

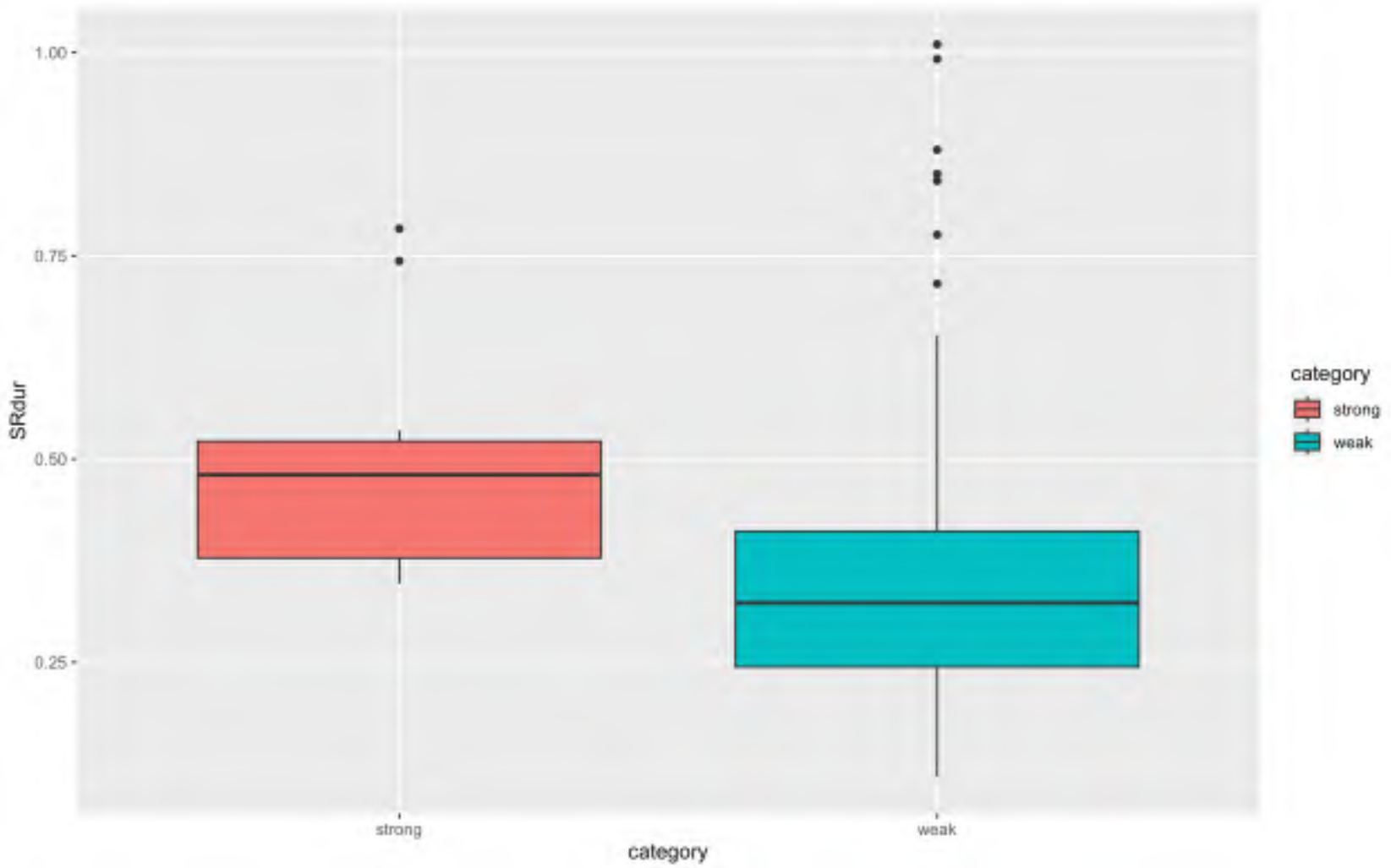
CORPUS

- A total 430 sentences with determiner + noun
- 161 with good audio quality



RESULTS

- The hypothesis is confirmed. Strong determiners (*die* or *dat*) have longer duration values in Dutch



THAT'S NOT ALL...

- During the qualitative inspection of the files, it was observed that within the weak determiner category there were extremely reduced forms (very short duration, with more centralized phonemes).

  aan a boeren, *in ə geschiedenis*, *vant coronavirus*

- Interestingly, these nouns commonly denote a globally unique entity or an abstract concept. This phenomenon has stayed under the radar until now!

ANOTHER EXAMPLE

(3) (*L'avite fatta nu poco tardi* 'You're running a bit late')
(*Eh! Lo saccio, ma c'aggia fà* 'Yeah, I know, but it's not my fault')

Chella *muglierema* *nun* *me* *vuleva*
that wife=my not me wanted

fà *venì*
make come

'My wife didn't want to let me out'

Ledgeway, Adam N. 2010. Subject licensing in CP: the Neapolitan double-subject Construction. In Paola Benincà & Nicola Munaro (eds.), *Mapping the left periphery*, 257–296. Oxford: Oxford University Press.

THE ADVANTAGES OF PYRLATO

- Pyrlato for ecologically valid, quick and convenient acoustic data collection and annotation
- Even web-based data like those from Pyrlato can be used to test hypotheses
- Data from spontaneous contexts present latent phenomena much more easily than in controlled laboratory-like situations.

⇝ Pyrlato allows us to make new observations and advance our understanding of natural language

ECOLOGY AND DIACHRONY

- Language variation ⇝ language change
- Apparent time (WuG) vs. historical data
- Dialects as frozen language history

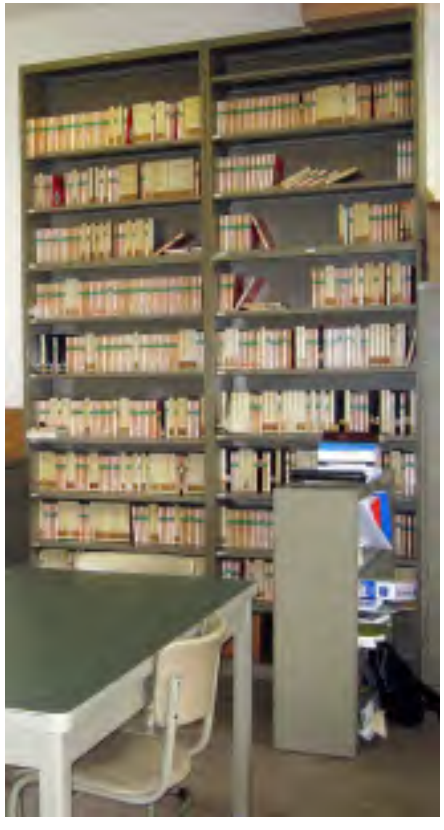
(4) Want ic ne wille niet, broeder, dat ghi onwetende sijt
'Because I do not want you to be unknowing, brother.'
(Middle Dutch, *Lectionarium Amsterdam* 1348)

(5) (To the cleaning lady after her surgery:)
K verwachten je de eerste weke nie vu te kusen. K=en gon=t wel zelve doen
'I don't expect you'll come in for cleaning the first week. I'll do it myself'
(I124p Lapscheure; L.Haegeman, p.c.)

GCND(+)

H116p	*Torhout*	veldwerker	[v=359] Met zo n weer je kun nie veel doen. [/v]	context		
		informant3	[a=n] Met zukke weer kun je nie veel doen. [/a] de drie informanten keuren deze zin af; nochtans komen er nogal wat inversieloze zinnen voor in de spontane spraak.	context		
N034p	*Hooglede*	veldwerker1	[v=359] Mee zulk een weer je kun nie veel doen ee.[/v]	context		
		informant1	[a=n] Me zuk n were kunje nie vele doen buitn.[/a] kun je	context		
		informant1	[a=n] Azo moet zijn.[/a] Hoewel in spontane spraak toch geregeld hoofdzinsorde is gevonden waar inversie wordt verwacht in AN, dus niet helemaal betrouwbaar, deze afwijzing?	context		

- Some (syntactic) structures that may be typical for dialects resist elicitation (hi, *chillo*!)
- Large archives of recodings of spontaneous (historical) dialect speech – *Stemmen uit het verleden*, *Nederlandse Dialectenbank*



GHENT UNIVERSITY

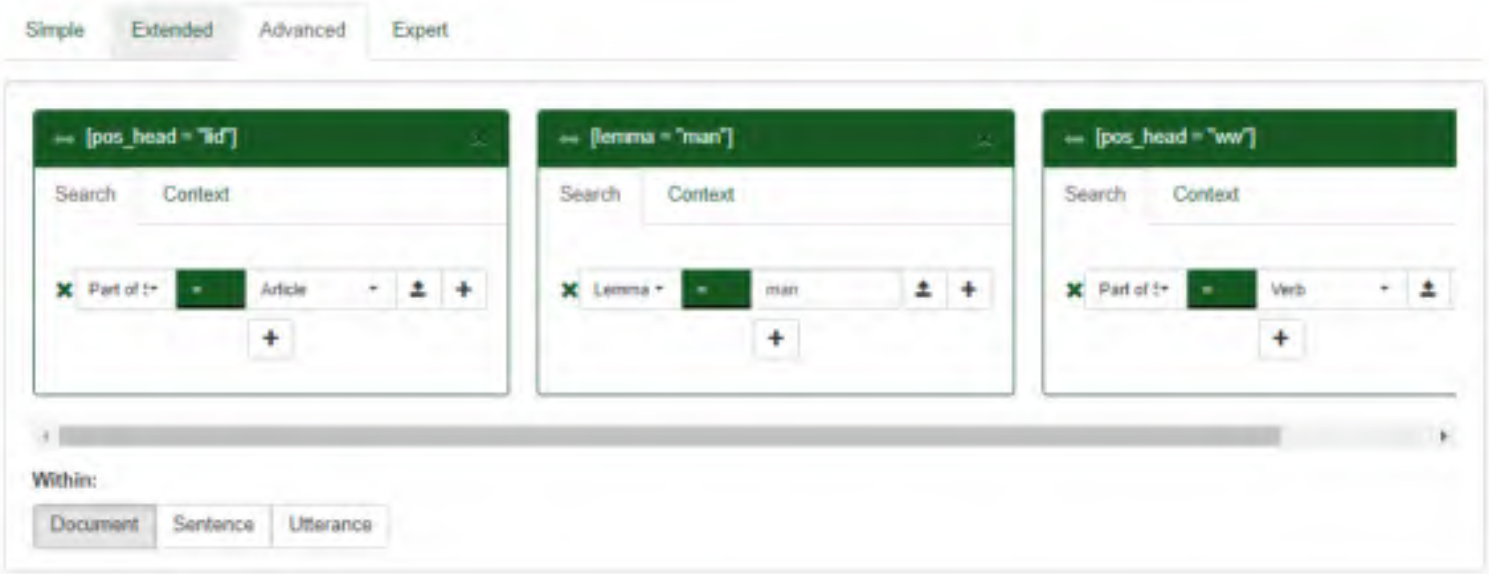
GCND(+)

- Transcription and annotation (POS, (dependency) parsing) in two FWO Medium Size Infrastructure projects (2020–2024; 2024–2028)
- so far 1206 speakers (oldest born 1871), ca. 500 hours of speech in 650 recordings from 639 places (to be expanded in GCND+!)

GCND

Gesproken Corpus van Nederlandse Dialecten

fwo



P041p_1: Vlaams-Brabant, Diest

...hè . eer dat aan **de man komt** . ja dat is altijd... de man komen lid(bep,stan,rest) n(soort,ev,basis,zijd,stan) ww(pv,tgw,met-t)

..zo een hoop en en dat geroddel ze komen juist bijeen voor wat te roddelen . dat is waar ze weten val alleman iets . daar moet ich niet van hebben . dan weet je ook veel nieuws . oh ja veel leugen veel bijgeroddeld hè . eer dat aan **de man komt** ja dat is altijd hè . hum . ja . het is misschien genoeg met z'n ??? we hadden moeten repeteren hè maar het zou beter geweest hebben . ja . ja ...

Property	value		
Normalised Transcription	de	man	komt
Dialect Transcription	de	man	komt
Lemma	de	man	komen
Part of Speech (full)	lid(bep,stan,rest)	n(soort,ev,basis,zijd,stan)	ww(pv,tgw,met-t)

WHAT ABOUT INTERPRETATION?

Summing up:

- Without natural / ecologically valid data, our interpretation of linguistic phenomena is biased
- Certain linguistic phenomena require specific discourse contexts and resist elicitation (e.g. inversionless V3 or mirative *en* in Southern Dutch dialects, expletive *chello/a, chillo* in Southern Italian dialects), or could not be elicited in the same quality and with the same distribution as in the "wild" (e.g. phonetic reduction of weak determiners)

⇒ "Wild" data may be harder to interpret because they are less controllable, but they form a valid —and in some cases indispensable— complementation to structured / controlled data

Thank you!

Claudia Crocco, Anne Breitbarth, Alexandra Simonenko, Kim Groothuis, Giuseppe Magistro, Giovanni Leo

Department of Linguistics
Research group DiaLing

E {claudia.crocco|anne.breitbarth|alexandra.simonenko|kim.groothuis|giuseppe.magistro|giovanni.leo}@ugent.be

www.ugent.be