FACULTEIT LETTEREN EN WIJSBEGEERTE

VAKGROEP TAALKUNDE / ΔIALING & GLIMS

Gauthier Delaby, Lien Hellebaut, Ulrike Vogl

Het DIRT-corpus (Dutch In Reality-Tv): Bron voor onderzoek naar informeel spontaan gesproken Nederlands

Wat?

- Het DIRT-corpus bestaat uit transcripties van informeel gesproken **Belgisch Nederlands** en **Nederlands Nederlands** uit reality-tv zoals **De Mol, Chateau Meiland, Temptation Island.** Het is verrijkt met metadata en bevat informatie over de regionale afkomst, gender, opleiding & leeftijd van de sprekers.
- **Transcripties** worden sinds 2021 gemaakt door (job)studenten Nederlandse taalkunde met behulp van een **transcriptieprotocol** gebaseerd op het protocol van het GCND (Ghyselen et al. 2020). DIRT wordt 2024-26 ondersteund door BOF Basisfinanciering.

Waarom?

- Het Nederlands in Vlaanderen wordt gekenmerkt door een situatie van **diaglossie** (Auer 2005), met een continuüm van de standaardtaal tot de dialecten, met daartussen een intermediaire vorm (*tussentaal*). Bestaande grote corpora omvatten niet alle taalvormen op dit continuüm:

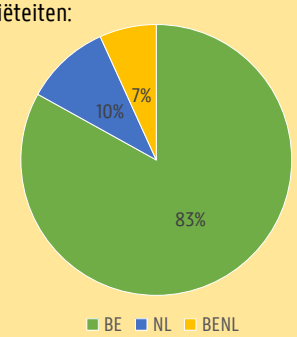
  ✓ *SoNaR-corpus*: geschreven tussentaal en standaardtaal
  ✓ *Gesproken Corpus van de Nederlandse Dialecten* (*GCND*): gesproken dialectische taal
  ✓ *Corpus Gesproken Nederlands (CGN)*: gesproken taal, formeel & informeel, maar eerder standaardtalig doordat sprekers de instructie kregen standaardtalig te spreken
  → **Nood aan een corpus met informeel gesproken Nederlands**

- Bestaand sociolinguïstisch onderzoek focust vaak op nationale variatie maar laat andere soorten sociale variatie buiten beschouwing doordat deze informatie slechts beperkt beschikbaar is in bestaande corpora.
  → **Nood aan een corpus met metadata over sprekers die de BN/NN-dichotomie overstijgt**

- Reality-tv vormt een interessante bron voor onderzoek naar informeel gesproken Nederlands:
  ✓ **Rijke sociale stratificatie** van de sprekers: sprekers van verschillende regio's, leeftijden en sociale achtergronden worden vertegenwoordigd (Zenner et al. 2009: 28)
  ✓ **Opnames in hoge kwaliteit**: mogelijkheid tot analyse van buitentalige context van bepaalde uitingen (Zenner & Van De Mieroop 2017: §2.2)
  ✓ **Jaarlijks veel nieuwe reality-programma's**: laat toe om een groeiend corpus te ontwikkelen met het oog op diachroon onderzoek
  ✓ Deelnemers worden langere tijd gevolgd, in **verschillende contexten en in situaties met veranderende sociale dynamieken**

Stand van zaken

✓ 55 afleveringen getranscribeerd
  - Uit 35 verschillende reality-programma's
  - 34 uur en 36 minuten spraak
✓ 286.718 woorden
✓ Verdeling van materiaal over nationale variëteiten:



✓ 33 van de 35 programma's uit de **periode 2017-2023**
✓ **644 sprekers**
  - Gemiddeld aantal woorden per spreker in het corpus: 438,38
  - 263 sprekers met meer dan 200 woorden in het corpus

Van bron naar begrip?

Is taalgebruik in reality-tv altijd spontaan?
❑ Deelnemers zijn zich bewust van de camera
❑ Sommige onderdelen (voice-over) zijn gescript

(Mogelijk) DIRT-onderzoek

❖ Vloeken Nederlanders anders dan wij?
❖ Vloeken jongeren vaker in het Engels?
❖ Wat bedoelen we met de interjecties *amai, allez* (Declercq 2022) en *ça va*?
❖ Zeggen vrouwen *uhm* en mannen *uh*?
❖ Is de groene of de rode woordvolgorde "informeler"?

Plannen voor de toekomst

(1) Oudere seizoenen van reality-tv (eind jaren 90 / begin 2000) transcriberen om ook diachroon taalkundig onderzoek op basis van DIRT mogelijk te maken

(2) Nederlandse en Surinaamse reality-programma's toevoegen om voor meer evenwicht te zorgen binnen het Nederlandse taalgebied

(3) Lemmatisering, woordsoortannotatie en parsing

(4) Gebruik van automatic speech recognition verkennen, ter aanvulling van manuele transcripties

Referenties

- Auer, P. (2005). Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. *Perspectives on variation: Sociolinguistic, historical, comparative*, 7, 7-42.
- Declercq, F. (2022). *'Allee hup'*: allee *in spontaan gesproken Nederlands - een semantisch-pragmatisch onderzoek naar de betekenissen en het gebruik van* allee *in Vlaams-Nederlandse reality-tv*. Masterproef, Universiteit Gent.
- Ghyselen, A., Van Keymeulen, J., Farasyn, M., Hellebaut, L., & Breitbarth, A. (2020). Het transcriptieprotocol van het Gesproken Corpus van de Nederlandse Dialecten (GCND). *Handelingen van de Koninklijke Commissie voor Toponymie & Dialectologie*, 92, 83–115.
- Zenner, E., Geeraerts, D., & Speelman, D. (2009). Expeditie Tussentaal: leeftijd, identiteit en context. *Nederlandse Taalkunde*, 14, 26-44.
- Zenner, E., & Van De Mieroop, D. (2017). The social and pragmatic function of English in weak contact situations: Ingroup and outgroup marking in the Dutch reality TV show Expeditie Robinson. *Journal of Pragmatics*, *113*, 77-88.

Contact

dirt@ugent.be
www.dirt.ugent.be

Universiteit Gent
@ugent
Ghent University