



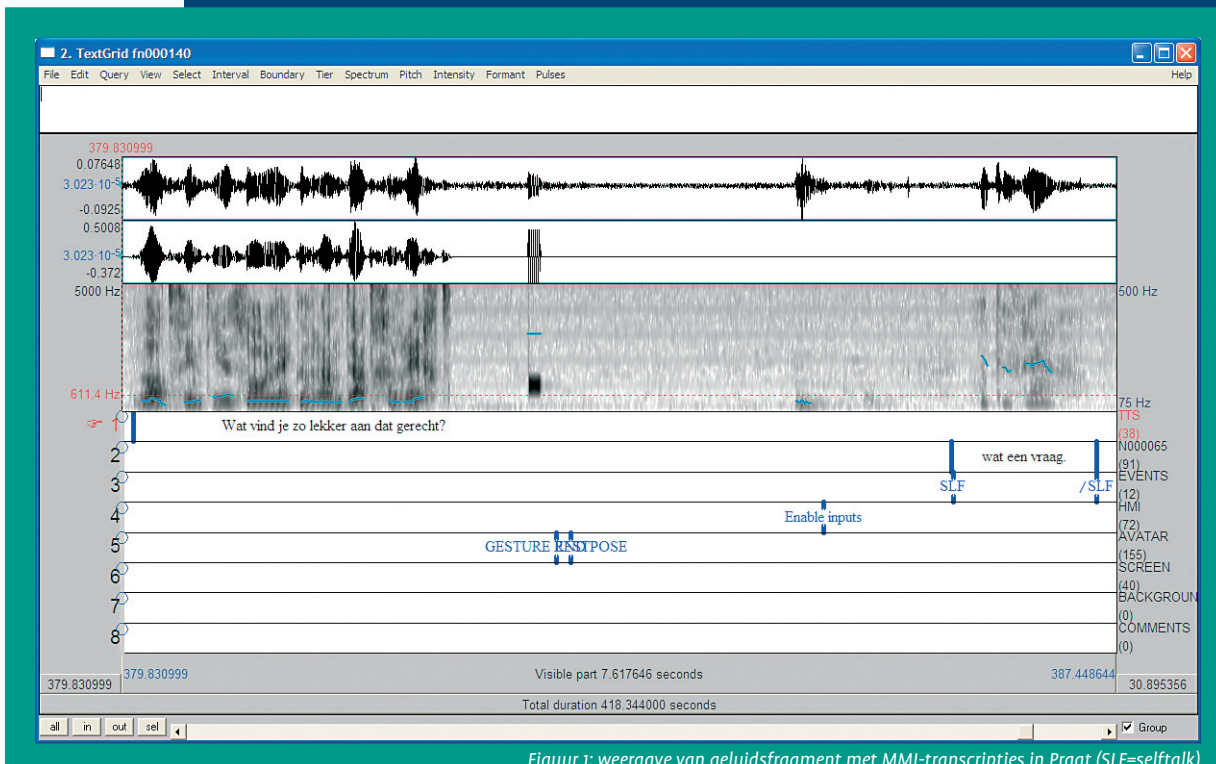
JASMIN-spraakcorpus

Wat is het JASMIN-spraakcorpus?

Het JASMIN-spraakcorpus is een verzameling van ca. 115 uur Nederlandse spraak van jongeren, anderstaligen en senioren, woonachtig in Vlaanderen en Nederland. De spraakopnames bestaan uit voorgelezen teksten en mens-machinedialogen, en zijn verrijkt met verschillende annotatie-lagen. Het JASMIN-spraakcorpus is een aanvulling op het Corpus Gesproken Nederlands (CGN), dat hedendaagse Nederlandse spraak van Vlamingen en Nederlanders bevat.

Het JASMIN-spraakcorpus omvat:

- **994 Spraakfragmenten** onderverdeeld in twee componenten; voorgelezen tekst en dialogen met een spraakcomputer (mens-machine-interactie). De data is georganiseerd als twee aanvullende componenten van het CGN (bestaande uit 15 componenten a t/m o); comp-p (dialogen) en comp-q (voorgelezen)
- **JASMIN-lexicon:** alle woorden (en woordvormen) die voorkomen in het spraakmateriaal, verdeeld in twee alfabetisch geordende lijsten van het Nederlands en Vlaams. Alle woorden zijn daarnaast ook voorzien van een fonetische transcriptie (txt)
- **Metadata:** informatie over de sprekers, de opnames en de data (txt en xls)



Figuur 1: weergave van geluidsfragment met MMI-transcripties in Praat (SLF=selftalk)



JASMIN-spraakcorpus

Alle spraakfragmenten zijn voorzien van de volgende annotaties:

- Handmatige **orthografische transcriptie** (uitgeschreven spraak) volgens de richtlijnen van het CGN
- Automatisch gegenereerde **fonetische transcripties** (klankweergave van wat er gezegd werd)
- **MMI-transcripties** van events die typerend zijn voor mens-machinedialogen (in zichzelf praten, herhaling, hyperarticulatie, stemverheffing etc.)
- Automatisch gegenereerde **part-of-speechtags** (morfologische en woordsoortinformatie)

Documentatie en software

Corpusdocumentatie en de transcriptieprotocollen worden meegeleverd met het corpus. Waar nodig wordt verwezen naar publicaties en (CGN-)protocollen. De spraakbestanden kunnen samen met de annotaties bekeken worden in het programma Praat. Meer informatie vindt u op www.inl.nl/producten, Productcatalogus, Corpora, JASMIN-spraakcorpus.

Distributie en prijzen

Het JASMIN-spraakcorpus wordt door de TST-Centrale gedistribueerd op 1 cd (documentatie) en 6 dvd's (geluidsbestanden met annotaties), of op externe harddisk. Het totale corpus is ± 23 GB groot.

prijs commercieel*: op aanvraag

prijs niet-commercieel*: € 0,-

**de prijzen zijn exclusief btw en eventuele verzend- en handlingkosten*

Het JASMIN-spraakcorpus is het resultaat van het JASMIN-CGN-project dat werd gefinancierd door het STEVIN-programma van de Nederlandse Taalunie. Het corpus is tot stand gekomen door een samenwerking tussen CLST - RU Nijmegen (projectcoördinator), ESAT - KU Leuven en TalkingHome. De rechten op het corpus zijn in handen van de Nederlandse Taalunie.

Colofon

Deze Product Sheet is een uitgave van de Centrale voor Taal- en Spraaktechnologie (TST-Centrale), een initiatief van de Nederlandse Taalunie (NTU) ondergebracht bij het Instituut voor Nederlandse Lexicologie (INL). De TST-Centrale wordt gefinancierd door de NTU.

E-mail: servicedesk@inl.nl

Ontwerp: Swantje Haage Ontwerp, Amsterdam