

Automata, Too

Henk van den Heuvel

CLST, Radboud Universiteit Nijmegen

Αὐτονομὰτὰ Τε

Het project

- Opvolger van Autonomata (2005-2007)
- Gesubsidieerd in de 3e open call van STEVIN
- Toepassingsgericht project
- Start: 1 februari 2008
- Einde: 1 februari 2010
- Volume: € 416,750

AUTONOMATA Too

- CLST, Radboud Universiteit Nijmegen (coordinator): Henk van den Heuvel
- ELIS, Universiteit van Gent: Jean-Pierre Martens
- Nuance: Bart d'Hoore
- TeleAtlas: Luc Peirlinckx, Luc Mortier
- UiL-OTS: Gerrit Bloothoof

letteren
Utrecht Institute of Linguistics OTS
letteren
Faculty of Arts
letteren *letteren* *letteren*

ELIS
DSSP


NUANCE



- Hetzelfde consortium als in Autonomata

ru | STEVIN

Doelen van het project

- *ASR van POIs verbeteren*
- *Bouwen van demo-applicatie om proof of concept te laten zien*

Achtergrond van het project

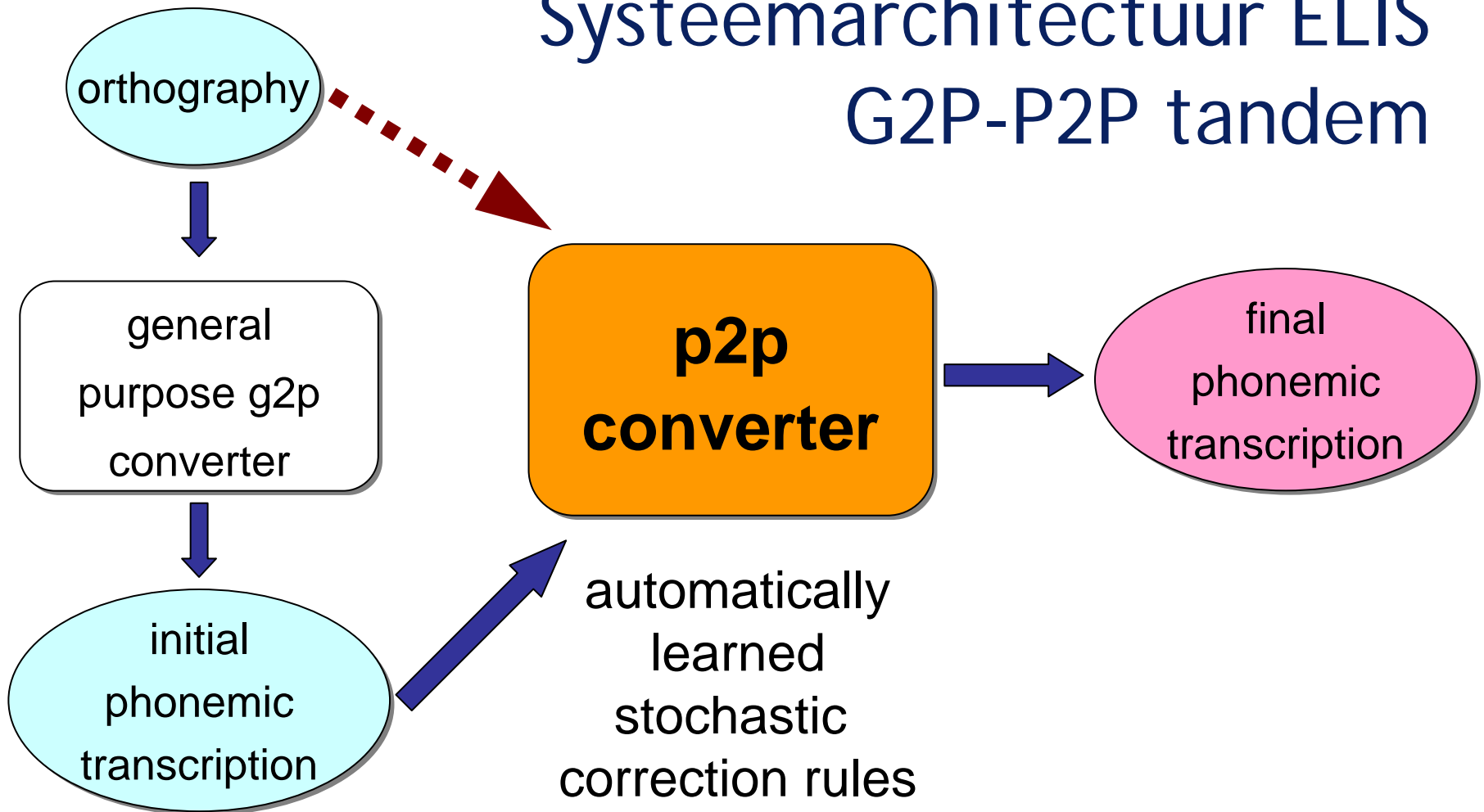
Wat zijn de specifieke problemen met ASR van namen?

- G2Ps voor gewone woorden werken niet goed voor namen vanwege:
 - Gefossilizeerde spellingen
 - Buitenlandse origine van namen
- Allerlei (inter)culturele verschijnselen veroorzaken veel varianten in de uitspraak van namen (fonemen / woordklemtoon):
 - NL/VL-sprekers die NL/VL namen uitspreken
 - NL/VL-sprekers die buitenlandse namen uitspreken
 - Anderstaligen die NL/VL namen uitspreken

Wat zijn de resultaten van Autonomata (I)?

1. P2P leersoftware en specifieke P2Ps om G2P-omzetting te verbeteren
2. Corpus met gesproken namen

Systemarchitectuur ELIS G2P-P2P tandem



Automata, Transfer Of Output

Demonstrator: Horeca in twee steden

- Andere typen namen dan in Automata:
 - P2P leersoftware om verbeterde foneemtranscripties te maken
- Reële uitspraken in herkenner:
 - D.m.v. modellering van patronen die worden gevonden in het namencorpus
- Inachtneming van interculturele aspecten:
 - Varianten binnen de Nederlandse foneemset
 - Varianten buiten de Nederlandse foneemset

AUTONOMATA Too

Stand van zaken

- WP1 (TeleAtlas): Dataselectie
 - Fon getranscribeerde straatnamen voor UK en FR geleverd
 - Fon. getranscribeerde POI entries NL en UK geleverd
 - Fon. getranscribeerde POI entries VL en FR in oktober verwacht
- WP2 (Nuance): ASR & prototype
 - VOCON 3.0 versie geleverd
 - Eerste versie prototype in september verwacht
- WP3 (ELIS): Uitbreidingen Autonomata
 - G2Ps voor Engels, Frans & Duits geleverd
 - Aanpassingen Autonomata G2P toolbox geleverd
- WP4 (CLST): Selectie mono-linguale uitspraakvarianten
 - Onderzoek naar P2Ps voor uitspraakvarianten van namen loopt
- WP5 (ELIS): Selectie multi-linguale uitspraakvarianten
 - In opstartfase (werknemer aangetrokken)
- WP6 (UiL-OTS): Evaluatie technologie en demonstrator
 - Design voor datacollectie/-transcriptie in november verwacht

AUTONOMATA Too

dANK vor y andAxt

WP1 (TeleAtlas): Data selectie & voorbereiding

- Data uit Autonomata:
 - Straat- en persoonsnamen (NL, VL)
- Nieuw in Autonomata Too:
 - Straatnamen: UK, FR
 - 20.000 fon. transcripties per taal
 - POIs (Accommodation/Food&Beverages): NL, VL, UK, FR
 - 5.000 fon. transcripties per taal

WP2 (Nuance): ASR technologie & prototype

- VOCON3200 spraakherkenner
 1. basisversie
 2. met multi-linguale AMs
- Baseline prototype van de demonstrator
- Finaal prototype van de demonstrator

WP3 (ELIS):

Uitbreidingen op Autonomata

- G2Ps voor Engels, Frans & Duits (Nuance)
- P2Ps voor alle naamsoorten behandeld in project (steeds mapping naar de Nederlandse foneemset)
- Aanpassing van Autonomata software toolbox:
 - taalkeuze uitbreiden
 - vrije keuze van de foneemset (+ mapping vanuit LH+) toelaten
 - meerdere p2p's in cascade kunnen toepassen
- Gericht op verbeterde canonieke transcripties

WP4 (CLST):

Selectie van monolinguale uitspraakvarianten

1. Onderzoek naar patronen in de verschillen tussen canonieke en reële uitspraken (uit het naamcorpus)
 - M.b.v. P2P leertechnologie
2. Selectie van de meest succesrijke varianten voor het lexicon van de herkenner
 - M.b.v. akoestische feedbackloop in herkenner

WP5 (ELIS): Selectie van multilinguale uitspraakvarianten

- Onderzoek naar buitenlandse fonemen die best in het lexicon en in de AM-set aanwezig zijn, voor verbetering van:
 - De herkenning van buitenlandse namen gesproken door NL/VL sprekers
 - De herkenning van NL/VL namen gesproken door buitenlandse sprekers

WP6 (UiL-OTS):

Evaluatie van de technologie en de demonstrator

- Collectie opnamen van 40 NL/VL sprekers en 40 buitenlandse sprekers
 - Met de baseline versie van de demonstrator
- Annotatie van de data zoals in Autonomata
- Test met 40 NL/VL sprekers en 40 buitenlandse sprekers
 - Met de uiteindelijke versie van de demonstrator

AUTONOMATA Too

Status:

- Consortium agreement ondertekend, incl.
 - Licenses Nuance voor
 - VOCON-herkenner
 - G2Ps Engels, Frans, Duits
 - License TeleAtlas voor
 - POIs met fon. transcripties NL, VL, UK, FR
 - Straat- en persoonsnamen met fon. transcripties
- Personeel compleet

Automata, Transfer Of Output

- Resultaten van Automata:
 1. P2P-omzetter voor persoonsnamen: NL & VL
 2. P2P-omzetter voor plaats- en straatnamen: NL & VL
 3. P2P leersoftware, toepasbaar op:
 1. Andere typen namen
 2. Andere talen
 3. Niet canonieke uitspraken
 4. Transcriptietools die p2p's kunnen aanwenden

Autonomata Naamcorpus

120 uit Nederland	60 autochtoon	15 Noord- en Zuid-Holland
		15 Gelderland
		15 Groningen, Friesland, Drenthe
		15 Noord-Brabant, Limburg
	60 allochtoon	20 Engels
		20 Frans
		20 Marokkaans Arabisch
120 uit Vlaanderen	60 autochtoon	15 Brabants
		15 Oost-Vlaams
		15 West-Vlaams
		15 Limburgs
	60 allochtoon	20 Engels
		20 Frans
		20 Marokkaans Arabisch

Automata naamcorpus

Spraakmateriaal:

- 70% Nederlands/Vlaamse namen
- 10% Engelse namen
- 10% Franse/Turkse namen
- 10% Marokaanse namen

WP1 (TeleAtlas): Data selection & preparation

- Selectie van POIs voor toepassing, e.g.:
 - Official name: Sorrentos Italian Pizzeria & Hotel BVBA
 - Address: Wetstraat 16, 1000 Brussel
 - Classification: Restaurant
 - Brand/TStyle: Sorrentos
 - Cuisine Type:
 - Italian
 - Pizzeria
 - Company Type: BVBA
 - 2nd Classification: Hotel with Eating Place