



Sociolectometry and Pluricentric Languages

Dirk Geeraerts



KU Leuven

OE Kwantitatieve Lexicologie en Variationele Linguïstiek



or

The Dutches of the Low Countries

A Tale of Attraction and Repulsion

Starting-point

if we define a **pluricentric language**, in the sense of Cline (1992), as a language with different standard varieties and different geographically distinct sociolinguistic clines between the standard stratum and the dialect stratum

then we need a **sociolectometric approach** to study pluricentricity, i.e. an approach that quantifies the synchronic (dis)similarity and the diachronic convergence/divergence between varieties

Purpose

- give an overview of the sociolectometrical techniques developed and used by the QLVL research team in Leuven
- illustrate their application to Dutch as a pluricentric language

(handout)

Overview

- **background**: the Dutch situation
- **sociolectometrical basics**: the 1999 **Convergentie en divergentie in the Nederlandse woordenschat** study
- **challenges** for the sociolectometrical perspective:
 - 1° getting a better grip on meaning
 - 2° developing a multivariate methodology
 - 3° comparing subjective and objective distances

Part the First

In which an attentive audience is introduced
to the fascinating characters of our narrative,
familarly referred to as The Dutches

Terminological background

Belgian Dutch / Netherlandic Dutch

the native varieties of Dutch used in Belgium : The Netherlands

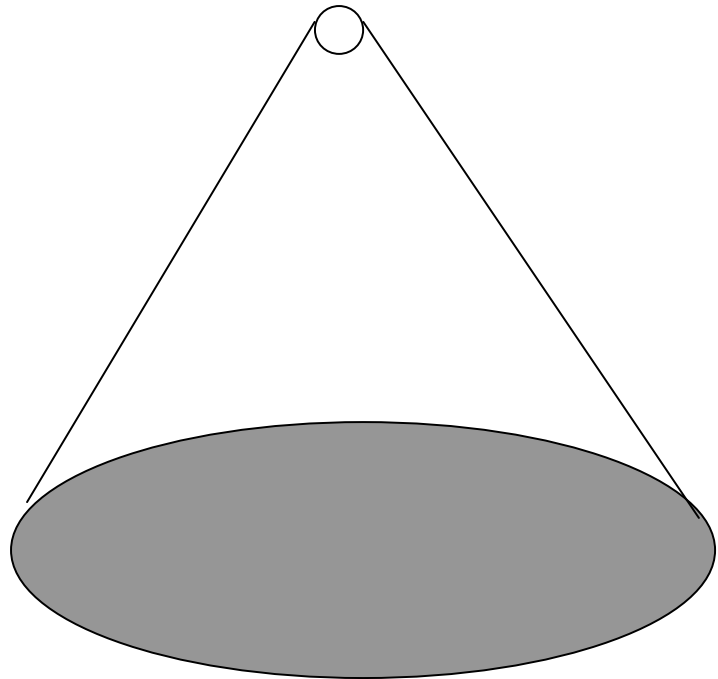
Standard Belgian / Netherlandic Dutch

the normative variety of Belgian Dutch

Colloquial Belgian / Netherlandic Dutch

the intermediate varieties of Belgian Dutch

Background: A model of variation

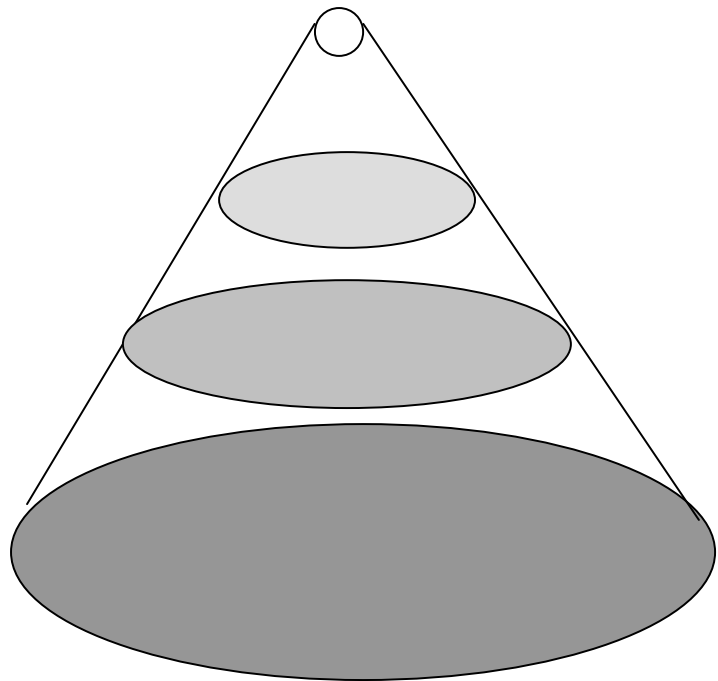


← supra-regional standardized Dutch

← local dialects

cp. Auer 2005

Background: A model of variation



← supra-regional standardized Dutch

← intermediate levels
(colloquial speech, regiolects ...)

← local dialects

cp. Auer 2005

Historical background

late standardization of Flanders

in The Netherlands since 17th century;

in Flanders (due to the Belgian context) slowly since late 19th century, as part of a nationalist movement,

and massively, as part of a modernization movement, after WWII

non-spontaneous standardization

top-down norm adoption instead of bottom-up norm emergence;

a pedagogical linguistic policy

Standard Belgian Dutch

pronunciation: features

no devoicing of voiced fricatives v, g, z

no hypercorrect voicing s > z (Zentraal Station)

no diphthongization of long monophthongs ee, oo

no lowering of diphthongs ij > aai, ou > au

no retroflex r (versus apical or uvular r)

less vowel reduction in unaccented syllables

Standard Belgian Dutch

grammar

some conspicuous differences

1 syntax: intrusion into clause final verb clusters

of verbal elements: ... dat hij de onderhandelingen zou gevoerd
hebben / zou hebben gevoerd

of nominal elements: de huurder zal op het huis waarin hij woont,
kunnen een bod doen / een bod kunnen doen

Standard Belgian Dutch

grammar

2 morphology: stronger tendency to maintain feminine gender
(even though the shift towards masculine gender is not absent)

de tafel, ze is gedekt

de kat, ze is spoorloos

Standard Belgian Dutch

lexicon

words that are generally accepted in Belgian Dutch but that are not (or hardly) known in The Netherlands

bissen 'een jaar overdoen' [repeat the year]

confituur 'jam' [jam]

croque-monsieur 'tosti' [toasted ham and cheese sandwich]

daim 'suède' [suede]

dominostekker 'verdeelstekker' [adapter]

fruitsap 'vruchtensap' [fruit juice]

Standard Belgian Dutch

lexicon

these words are to be distinguished from

1 words that refer to specifically Belgian institutions

schepen ‘± wethouder’ [alderman]

gouverneur ‘± commissaris van de koningin’ [provincial governor]

leefloon ‘uitkering van het OCMW aan personen die geen inkomen hebben’ [welfare cheque]

Standard Belgian Dutch

lexicon

2 words that typically belong to colloquial Belgian Dutch:

ambetant ‘vervelend’ [boring, annoying]

camion ‘vrachtwagen’ [truck]

camionette ‘bestelwagen’ [van]

bareel ‘slagboom’ [barrier, gate]

botten ‘laarzen’ [boots]

Colloquial Belgian Dutch

pronunciation

- deletion of final t or d in high frequency (function) words: **nie[t]**, **da[t]**, **wa[t]**, **goe[d]** ...
- deletion of initial h before vowel: **[h]eel**, **[h]uis**, **[h]ard** ...

in addition, CBD pronunciation is regionally colored

Colloquial Belgian Dutch

grammar

1 adnominal flektion: determiners, demonstrative pronouns, possessive pronouns, adjectives receive specific flektion depending on the gender of the noun (and the initial vowel)

Colloquial Belgian Dutch

grammar

1 adnominal flexion with masculine nouns

een stoel

ne stoel

geen stoel

gene stoel

die stoel

dieë stoel

de stoel

de stoel

een grote stoel

ne grote stoel

Colloquial Belgian Dutch

grammar

1 adnominal flexion with masculine nouns before h, b, d, t, vowel

een hond

nen hond

geen hond

genen hond

die hond

dieën hond

de hond

den hond

een grote hond

ne groten hond

een huilende hond

nen huilenden hond

Colloquial Belgian Dutch

grammar

1 adnominal flektion in definite NPs with neuter nouns:

loss of adjectival flektion

het witte paard

het wit paard

Colloquial Belgian Dutch

grammar

2 second person pronouns and verbal forms:

no distinction between T/V forms, but a single **gij**-system

with specific forms for inversion:

gaat jij / gade jij / gade

Colloquial Belgian Dutch

grammar

3 personal pronouns and reflexive pronouns:

ekik versus ik (inversion)

hem versus hij (inversion)

z'n eigen versus zichzelf

Colloquial Belgian Dutch

grammar

4 diminutives:

-ke, -ske, -eke versus -je, -tje, -pje

stoeleke, appelke, bakske

Colloquial Belgian Dutch

lexicon

as mentioned above: **ambetant**, **ambetanterik**, **camion**, **camionette**,
bareel, **botten** etc.

Part the Second

In which our narrative, with graphical detail,
unfolds the intricate and intriguing relationship
among The Dutches



Comparing sociolinguistic continua

a characteristic difference between Belgian Dutch and Netherlandic Dutch is the position of the colloquial varieties on the stratificational continuum between standard variety and local dialects (as revealed by a sociolectometrical approach); cp. Augusto Silva's work on European and Brazilian Portuguese

a lexical case study - why lexical?

ensures comparability of variables (markers of informality are different in SBD and CND)

A lexical case study

method

1) “onomasiological profile”

= a set of synonyms distinguished by their relative frequencies

jasje 2115 ; colbert 945 ; vest 1440

jasje 47% ; colbert 21% ; vest 32%

2) “uniformity”

the degree of overlap between the onomasiological profiles for a given concept in different language varieties

A lexical case study

materials

1950 - 1970 –1990

ca. 40.000 observations

two lexical fields: clothing, football

[scheidsrechter / referee / ref], [doelman / keeper / doelwachter],

[broek / pantalon], [jeans / spijkerbroek]

sources: nation-wide newspapers and magazines;

local shop windows (Leuven / Kortrijk; Leiden / Maastricht)

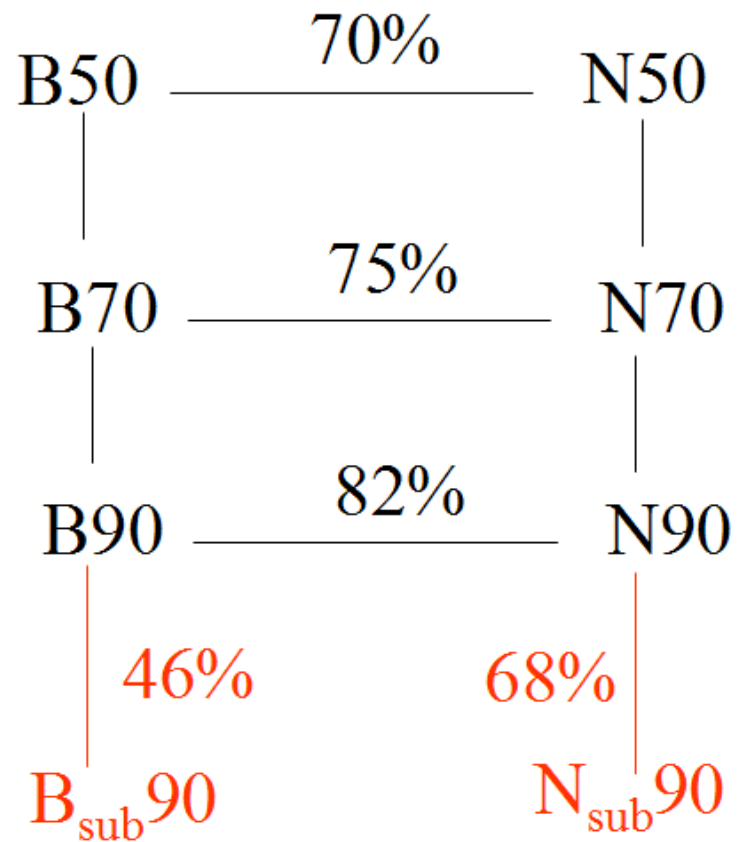
A lexical case study

results

B50/N50:	69.84
B70/N70:	74.59
B90/N90:	81.70
B90/B _{etal} 90:	45.90
N90/N _{etal} 90:	67.75

A lexical case study

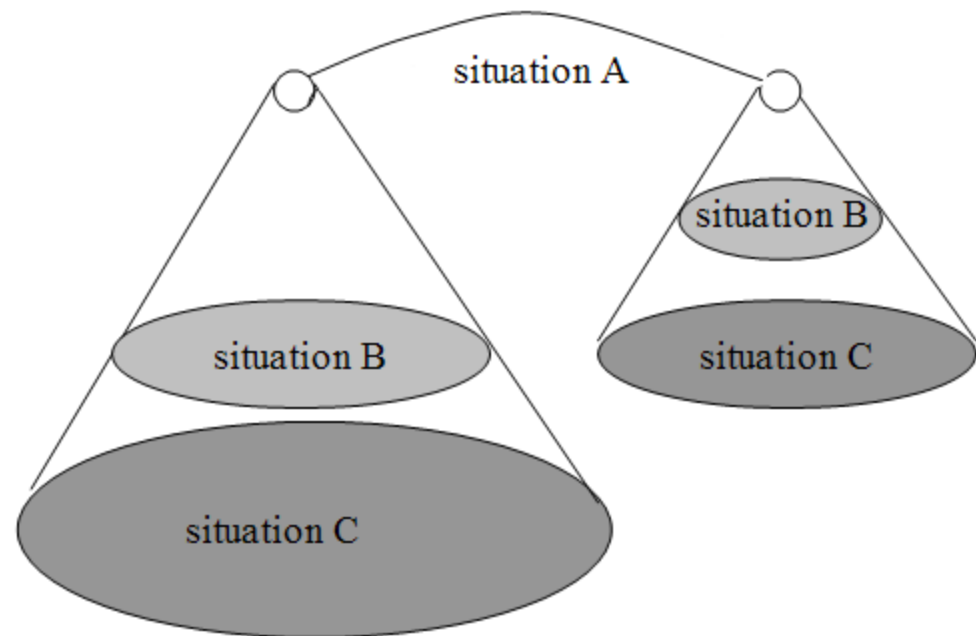
results



Comparing sociolinguistic continua

conclusion:

the stratificational continua
are different in the two
national varieties of Dutch
(consistently replicated)



Comparing sociolinguistic continua

Dutch is a pluricentric language

because Netherlandic Dutch and Belgian Dutch exhibit separate, different stratificational structures

but the relationship between the centers is asymmetrical

because there is (or has been) a normative orientation of Belgian Dutch with regard to Netherlandic Dutch

Part the Third

In which, to the bewilderment of the reader,
the plot of our tale thickens
in numerous numerical ways

Extensions of the basic approach

- sociolectometry assumes identity of meaning: how obvious is semantic equivalence?
- the comparison assumes that the lectal stratification is two-dimensional at most: what if we take a multivariate approach to lectal structure?
- the measurements are based on objective measures of language use only: what if we add a subjective dimension?

Extensions of the basic approach

- sociolectometry assumes identity of meaning: how obvious is semantic equivalence?
- the comparison assumes that the lectal stratification is two-dimensional at most: what if we take a multivariate approach to lectal structure?
- the measurements are based on objective measures of language use only: what if we add a subjective dimension?

Semantic equivalence

sociolectometrical distance calculations of the type mentioned depend on semantic equivalence: we can measure differences in regional etc. use of words because the words are semantically equivalent – but then how do we establish equivalence?

the problem is not restricted to words: it's a general problem for sociolinguistics in the Labovian sense

cp. the Labov – Lavandera debate:

Semantic equivalence

Labov:

“Social and stylistic variation presuppose the option of saying ‘the same thing’ in several different ways: that is, the variants are identical in reference or truth value, but opposed in their social and/or stylistic significance”

(1972: 271)

Semantic equivalence

Lavandera:

“It is inadequate at the current state of sociolinguistic research to extend to other levels of analysis of variation the notion of sociolinguistic variable originally developed on the basis of phonological data. The quantitative studies of variation which deal with morphological, syntactic, and lexical alternation suffer from the lack of an articulated theory of meanings”

(1978: 171)

Semantic equivalence

→ a sociolectometrical approach needs to meet the challenge of semantic variation: if we take a usage-based approach to language variation seriously, we will need to develop models to disentangle functional and formal variation and their interaction

Semantic equivalence

to illustrate: Speelman, Grondelaers & Geeraerts 2008 study the choice of adjectives in SND and SBD

(method: Stable Lexical Marker analysis, a variant developed by Dirk Speelman of the keyword analysis)

differences derive from various sources:

Semantic equivalence

- 'profile-based' differences of the type studied above

volledig, heel vs gans

- topical differences referring to the national context

federaal, Franstalig in SBD

- differences of conceptual stance

more subjective expressions in SND: prettig,

ontzettend, geweldig, prima ...

Semantic equivalence

- 'profile-based' differences of the type studied above
talking about the same thing with different words
 - topical differences referring to the national context
talking about different things
 - differences of conceptual stance
talking about the same things with different
perspectives and construals
- linguistic differences are not only linguistic; they may also be cultural

Semantic equivalence

in the context of our QLVL research group, we currently pursue two methods for getting a better grip on the problem of semantic equivalence: we use

- vector-based word space models to identify sets of near-synonyms
- regression models for the micro-analysis of near-synonyms, in which we include both lectal and semantic features

Extensions of the basic approach

- sociolectometry assumes identity of meaning: how obvious is semantic equivalence?
- the comparison assumes that the lectal stratification is two-dimensional at most: what if we take a multivariate approach to lectal structure?
- the measurements are based on objective measures of language use only: what if we add a subjective dimension?

A multivariate approach

morphological markers of Colloquial Belgian Dutch, as studied in the PhD thesis of Koen Plevoets (2008)

the linguistic variables

a large collection of phenomena from the nominal and verbal morphology of Dutch

A multivariate approach

the lectal parameters

Spoken Dutch Corpus (10 mio tokens, 1/3 Belgian Dutch)

differentiation according to register: 15 stylistic components,

relevant dimensions (prepared vs. spontaneous speech; private vs.

public speech), province, generation, gender, education level,

occupation level

A multivariate approach

register components

- a: Spontaneous conversations ('face-to-face')
- b: Interviews with teachers of Dutch
- c: Spontaneous telephone dialogues (recorded via a switchboard)
- d: Spontaneous telephone dialogues (recorded on MD via a local interface)
- e: Simulated business negotiations
- f: Interviews/discussions/debates (broadcast)
- g: (political) Discussions/debates/meetings (non-broadcast)
- h: Lessons recorded in the classroom
- i: Live (sports) commentaries (broadcast)
- j: Newsreports/reportages (broadcast)
- k: News (broadcast)
- l: Commentaries/columns/reviews (broadcast)
- m: Ceremonious speeches/sermons
- n: Lectures/seminars
- o: Written text read aloud

A multivariate approach

province

- brab: Flanders, central region (Antwerpen and Vlaams-Brabant)
- ovl: Flanders, transitional region (Oost-Vlaanderen)
- wvl: Flanders, peripheral region 1 (West-Vlaanderen)
- lim: Flanders, peripheral region 2 (Limburg)

gender M/F

generation

- pre: the pre-war generation, born between 1910 and 1929
- sil: the 'silent' generation, born between 1930 and 1939
- pro: the protest generation, born between 1940 and 1954
- los: the 'lost' generation, born between 1955 and 1970
- pra: the pragmatic generation, born after 1970

A multivariate approach

educational level

hie (high); mid (middle); low (low)

occupation

- occA: occupation in higher management or government
- occB: occupation requiring higher education
- occC: employed on the teaching or research staff in a university or school
- occD: employed in an administrative office or a service organisation
- occE: occupation not requiring any specified level
- occF: self-employed
- occG: politicians
- occH: employed in the media, entertainment or artistic sector
- occI: student, trainee
- occJ: having no job

A multivariate approach

the method

Correspondence Analysis

- similar for categorical data to Principal Component Analysis for numerical data
- clustering shows the similarity in behavior between the linguistic variables as well as the lectal ones

A multivariate approach

→ 'lects' as homogeneous clusters of sociostylistic features and linguistic features do not exist; the concept of clear stratificational levels and categorical language varieties is an idealization

nevertheless, dimensions can be discovered

e.g. register: no stratification, but a continuum along two dimensions (horizontal: public vs. private; vertical, within public): hierarchical vs. equal)

A multivariate approach

→ a sociolectometrical approach needs to meet the challenge of a bottom-up methodology: if we take the complexity of language use seriously, we will need to develop types of analysis that can cope with that complexity; that implies doing analyses at an aggregate level that is not the usual one in sociolinguistic research

Extensions of the basic approach

- sociolectometry assumes identity of meaning: how obvious is semantic equivalence?
- the comparison assumes that the lectal stratification is two-dimensional at most: what if we take a multivariate approach to lectal structure?
- the measurements are based on objective measures of language use only: what if we add a subjective dimension?

A perceptual approach

subjective distances between language varieties basically involve two sets of phenomena:

- perception, recognition, identification
- attitudes, evaluation, stereotyping

an example of perceptual sociolectometry: Impe (2010) on the mutual intelligibility of national and regional varieties of Dutch

A perceptual approach

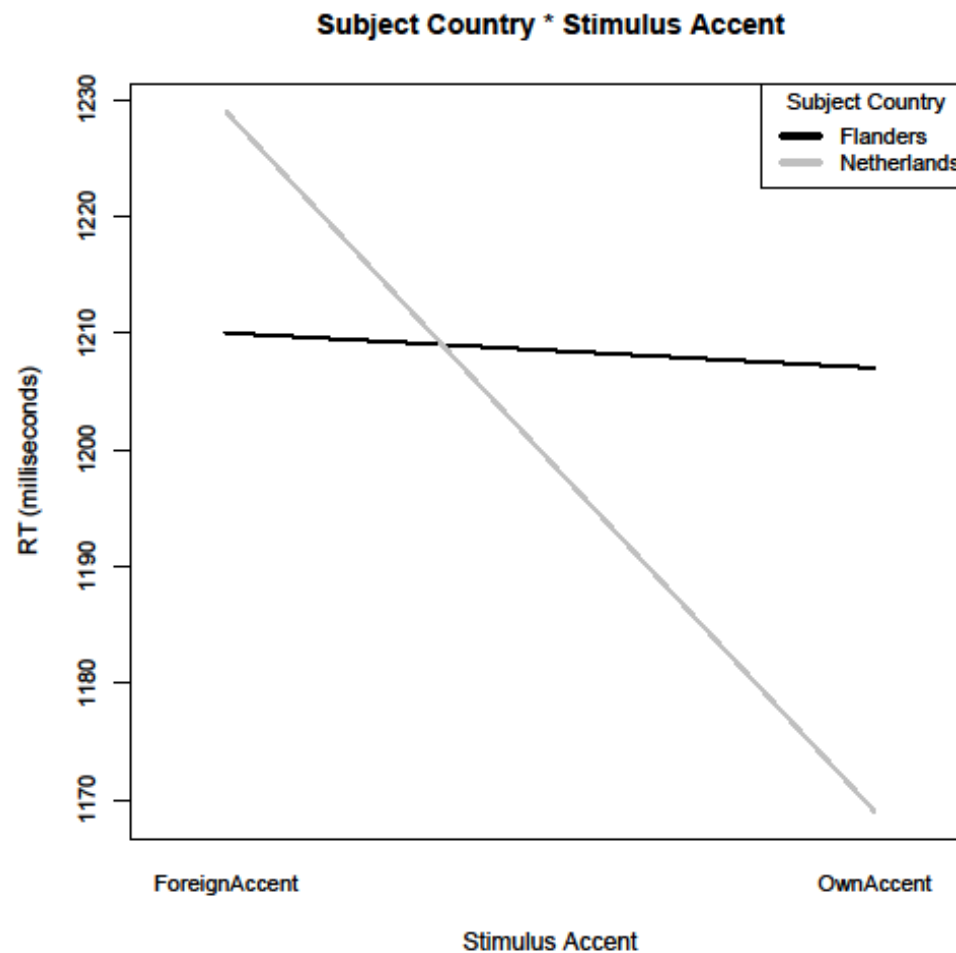
design of the study:

- auditory stimuli, consisting of real words and nonsense words, spoken with various regional and national accents
- reaction time experiment: “Is this a Dutch word?”
- tests subjects from 4 Belgian and 4 Netherlandic provinces

the results confirm the asymmetry of the relationship:

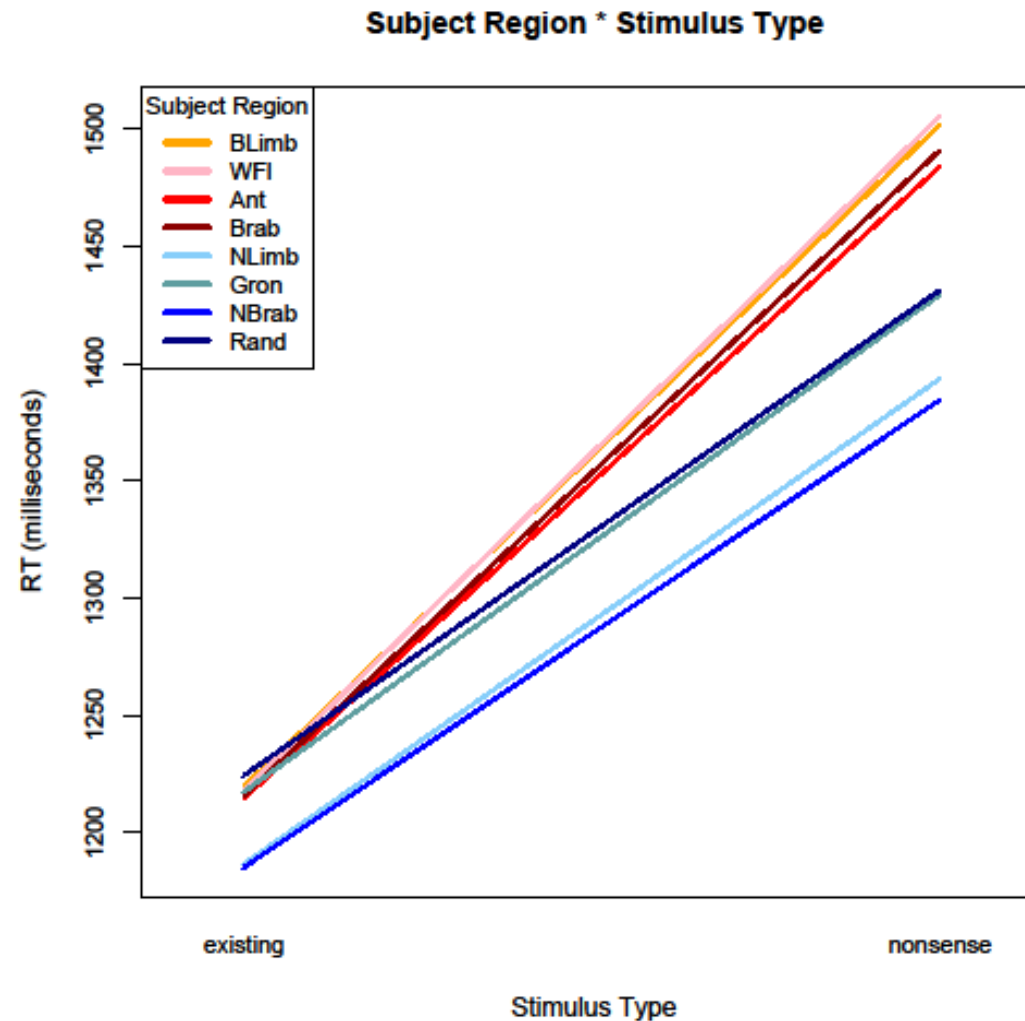
A perceptual approach

recognition of own national accent (existing words):



A perceptual approach

recognition of existing vs. nonsense words:



A perceptual approach

- speakers of Belgian Dutch are more attuned to the existing variety within Dutch than speakers of Netherlandic Dutch: they recognize existing variety more easily
- speakers of Belgian Dutch have less subjective confidence than speakers of Netherlandic Dutch: they reject non-existing words less easily

Part the un-Last

In which the benevolent but exhausted audience
is invited to the sequels in the
continuing story of The Dutches

Conclusions

- 1 Dutch in the Low Countries is a pluricentric language, but the relationship between the centers is both subjectively and objectively an asymmetrical one
- 2 sociolectometry is an important tool for the study of language-internal lectal variation in general, and the situation of pluricentric languages in particular

Conclusions

3 the methodology of sociolectometrical research will have to be further developed to meet the challenge of

- a bottom-up approach to multivariate linguistic diversity
- the interplay of objective and subjective distances
- the role of meaning and function in linguistic variation

Sequels

research into sociolectometrical methods in general and into the story of The Dutches in particular continues at QLVL;
further talks in the context of the present conference:

Thu 14.30 Dirk Speelman on controlling for meaning

Thu 15.00 Dirk De Hertog on applied lexical sociolectometry

Thu 15.30 Tom Ruetten on multivariate models

thank you !

for further information:

<http://wwwling.arts.kuleuven.be/qlvl>

dirk.geeraerts@arts.kuleuven.be

