FACULTY OF ARTS
AND PHILOSOPHY

Department of Linguistics
Research group DiaLing

TOWARDS BOTTOM-UP SYNTACTIC DIALECTOLOGY

Considerations on the use of parsed corpora of spontaneous dialect speech

Anne Breitbarth

REEDS, Amsterdam, 29–30 June 2023

GHENT
UNIVERSITY

OVERVIEW

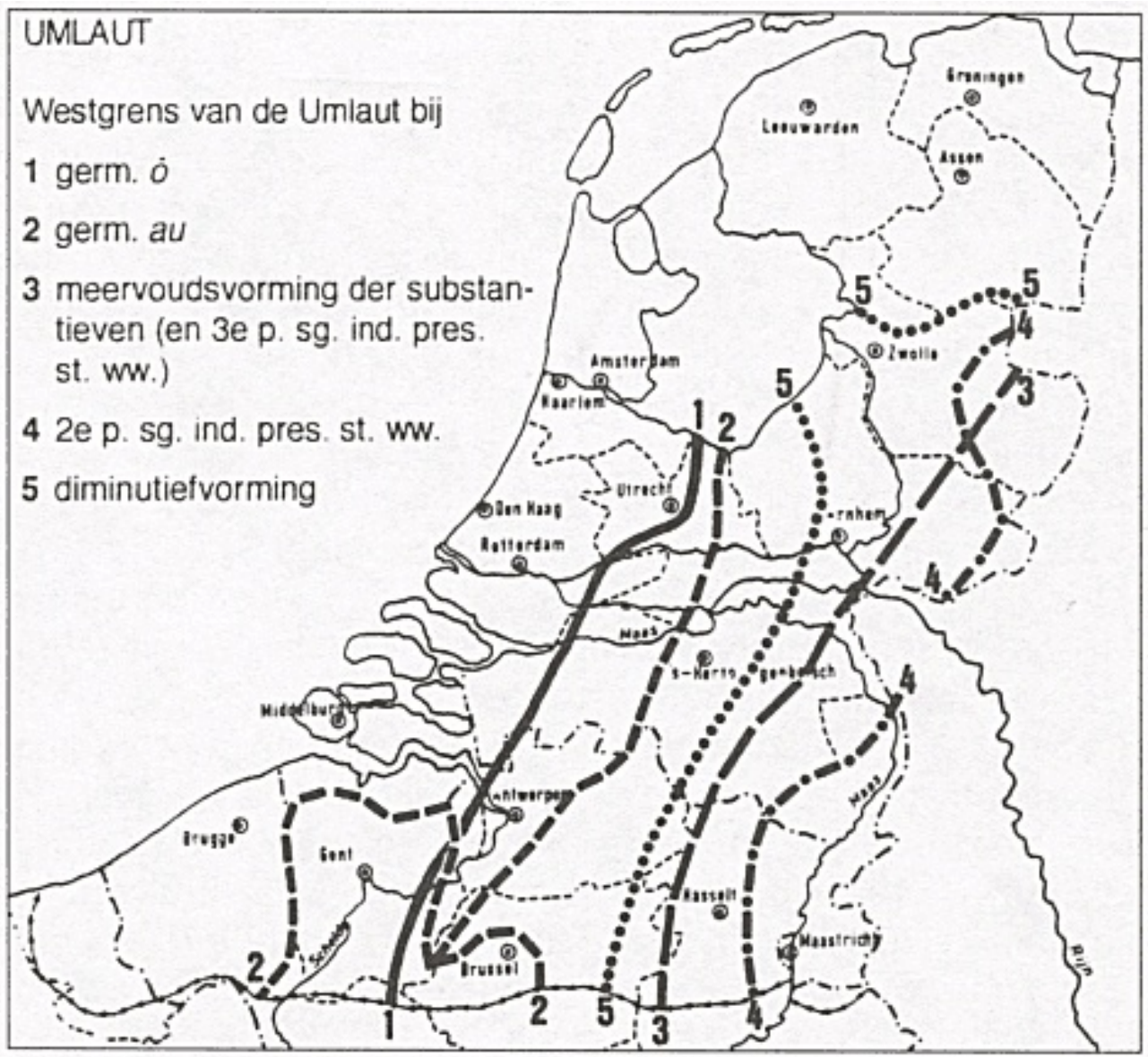
1. Top-down methods in dialectology
2. Bottom-up syntactic dialectology?
3. Outlook

TOP-DOWN METHODS IN DIALECTOLOGY

DIALECTOLOGY

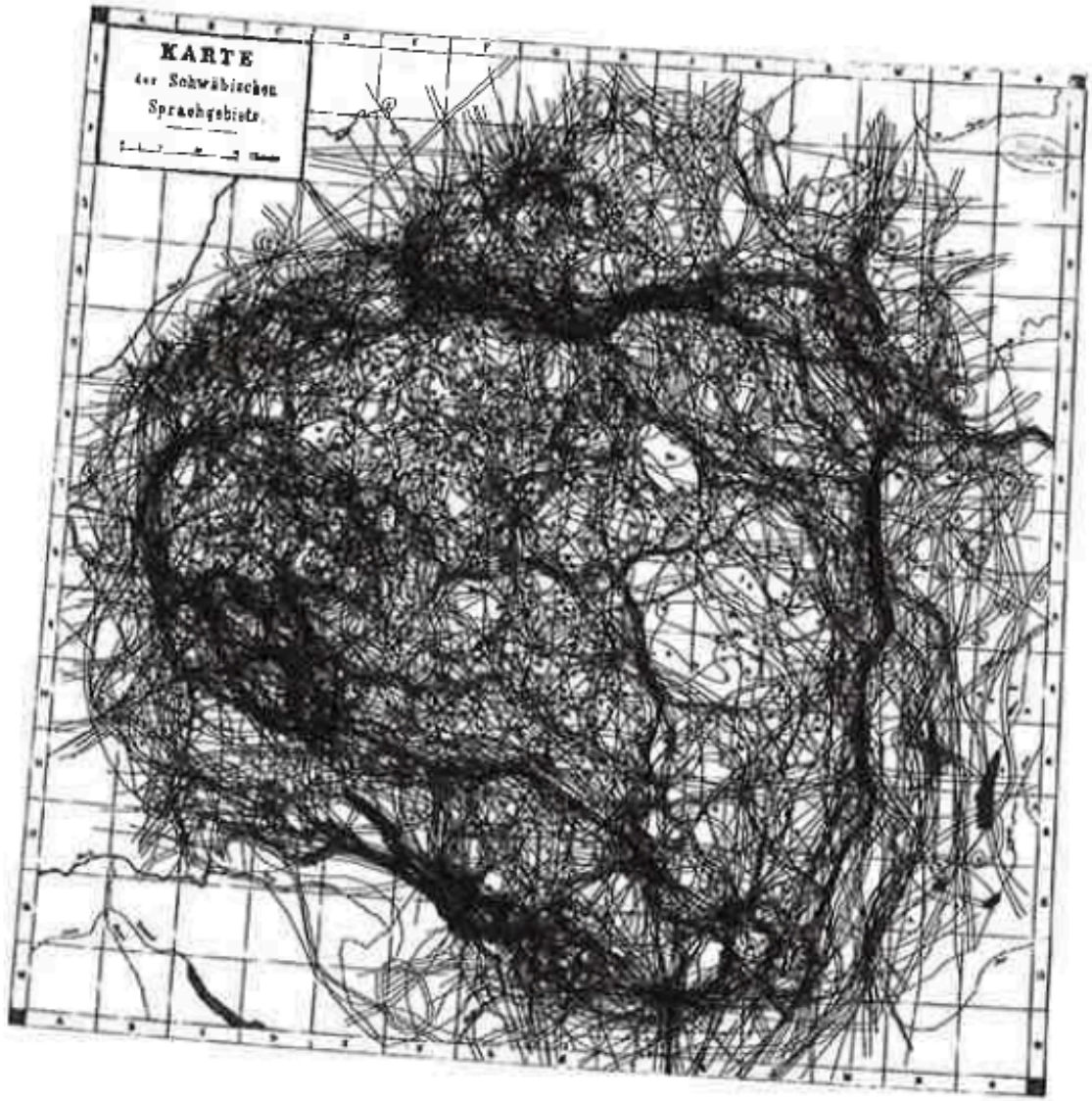
- Two main aims:
  1. Catalogue attested linguistic features per location
  2. Find spatial patterns in the variation between these features

DIALECTOLOGY



(from van Bree 1996: 232)

DIALECTOLOGY



(from Lameli 2013: 2)

DIALECTOLOGY

- Requirements:
  1. Naturalness
  2. Commensurability
- Consequence of this struggle to balance naturalness and commensurability: top-down approaches
  1. Elicited data (questionnaires, word lists, list of sentences, translation + judgment tasks)
  2. Pre-selected features for spoken corpora
- Advantages:
  1. Comparability
  2. Replicability
  3. Cost-effectiveness

TWO PROBLEMS

- Problems:
  1. Observer's Paradox: priming and accommodation
  2. Selection Bias (e.g. SAND: "Development of an inventory of existing knowledge on syntactic variation. Among other things, this included a study of the literature"; Barbiers et al. 2005:8)
- OP: well known, can also be shown to riddle syntactic dialect surveys (e.g. SAND)

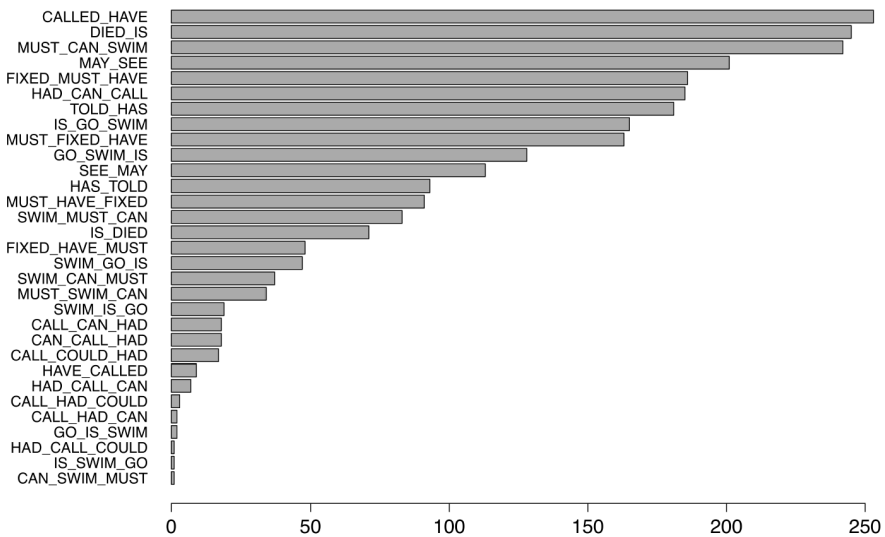


FIGURE 2. Frequency of the thirty-one verb cluster orders found in the SAND data.

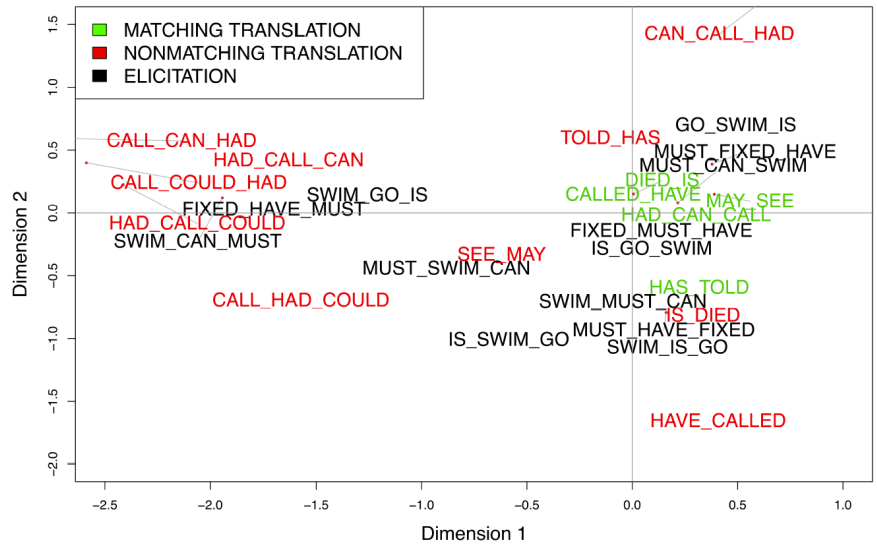


FIGURE 10. Two-dimensional representation of the SAND verb cluster data, color-coded according to question methodology.

(from Van Craenenbroeck et al. 2019)

TWO PROBLEMS

- Selection Bias:

ELUSIVE SYNTACTIC PHENOMENA

- Selection Bias is real: you only find what you search for (and sometimes not even that)

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		

(https://www.meertens.knaw.nl/sand/)

ELUSIVE SYNTACTIC PHENOMENA

- Elicitation can only find what was asked for –
- Certain phenomena, particularly such that are dependent of specific discourse contexts, have a tendency to be overlooked
- What about using recorded speech – Naturalness?

uniforme vragenlijst. Theoretisch krijgt men betere dialect-gegevens wanneer men er zich toe beperkt, spontane gesprekken af te luisteren en daaruit „in de vlucht" te noteren wat men kan ; men zou deze gesprekken, indien men met de nodige materiële middelen was voorzien, ook op band kunnen registreren en ze daarna in fonetische teksten omzetten. En ik wil graag erkennen dat een dergelijke documentatie zeer kostbaar zou zijn ter aanvulling en illustratie van het systematisch onderzoek met vaste vragenlijst.

(Blancquaert 1948: 12)

BOTTOM-UP SYNTACTIC DIALECTOLOGY?

EARLIER (BOTTOM-UP) PROPOSALS

- Szmrescanyi (2013): 57 pre-selected syntactic features based on earlier dialectological literature and atlases (most of it itself based on elicited data collection), relatively superficial (extracted semi-manually with scripts, corpus neither parts-of-speech tagged nor parsed) in a corpus of transcribed spontaneous dialect speech of English in the UK (FRED)
- Sanders (2010): bottom-up feature extraction from automatically parsed Swedish part of the ScanDiaSyn (Johannessen et al. 2014) corpus (the SweDiaSyn-part), 36,713 sentences from 49 locations > various automatic feature extraction methods, e.g. POS-trigrams, leaf-ancestor-paths, leaf-head-paths, PS-rules, ...
- Wolk (2014): bottom-up; POS-bigrams (not enough data for trigrams – sparsity!) from POS-tagged part of FRED (1 million out of 2,5 million tokens from 163 locations in 43 UK counties)

CHALLENGES FOR BOTTOM-UP APPROACHES

1. Comparability. Traditional dialectology works with word lists and sentence lists for a reason. Spontaneous speech/free text: What to compare?
   - More abstract syntactic patterns ✓
   - Frequencies ✓
   - Alternatives / Variants
2. Missing data: Accidentally unattested yet possible patterns
3. Linguistic theory: What's the relationship between (surface) patterns and underlying parameters?

SPONTANEOUS DIALECT SPEECH CORPORA

- parsed corpora of spontaneous dialect speech = rare
  - CORDIAL-SIN (Martins 2000-; Magro 2010) – ca. 600,000 word forms, 42 places
  - AAPCAppE (Tortora et al. 2017) – ca. 1 million word forms
- Both: Penn Treebank system
  - Constituency parsing (as opposed to dependency parsing) plus some dependency features (grammatical functions)
  - Relatively shallow trees:
    - no VP
    - all constituents of a clause are immediate daughters of IP
    - CP-layer only if e.g. wh-movement/relativisation/..

EXPERIMENT

- POS-trigrams (or even only bigrams) (as used by Nerbonne/Wiersma 2006, Sanders 2007;2010, Wolk 2014) may represent *some* syntactic information, but are still very close to the actual words (cf. Wolk 2014)
- Experiment: trigrams of constituent labels from the Penn-style syntactic annotation of e.g. CORDIAL-SIN
- Besides more word-level labels such as for verbs, the Penn scheme has phrasal-level annotations, and some information on grammatical function

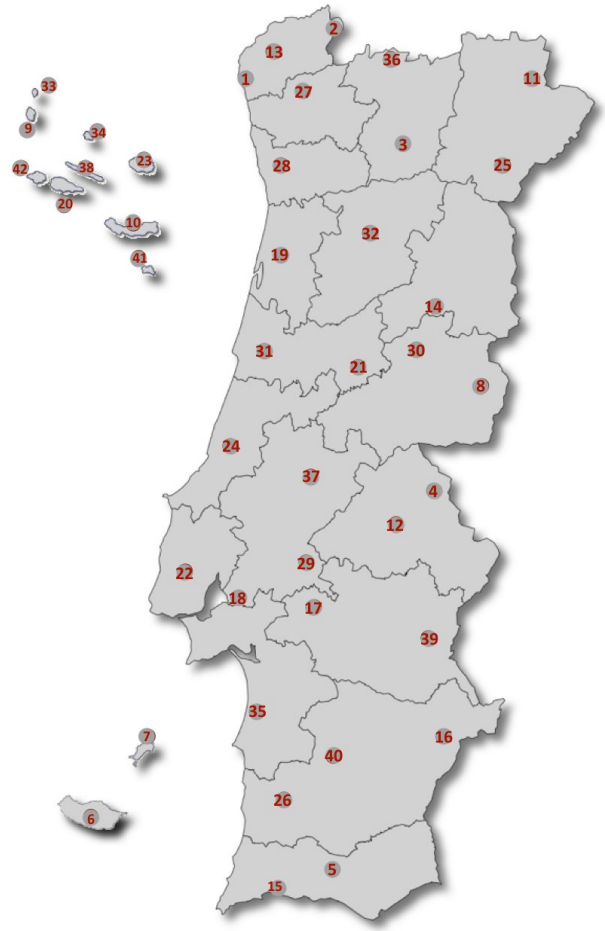
(1) *Nós*$_{\text{NP}-\text{LFD}}$ *dantes*$_{\text{ADVP}}$, [*@em @estas redes*]$_{\text{PP}}$, *era*$_{\text{SR}-\text{D}-\text{3S}}$ *raro*$_{\text{ADJP}-\text{PRD}}$ [[*o dia*] *que não se*
we before in these nets was rare the day that NEG REFL
pegava um, *dois lavagantes ou três*]$_{\text{CP}-\text{REL}}$]$_{\text{NP}-\text{SBJ}}$.
caught one two lobsters or three

(Vila Praia de Âncora, Adail *1925; INF-VPA01)

⇝ More sentence-level syntactic information

EXPERIMENT

1. Extracted all constituent labels (searching for all `IP-MAT`s) using CorpusSearch (Randall 2005) in the 42 places in the CORDIAL-SIN
2. Extracted all constituent label trigrams per place (unifying finite verb tags), padding sentence boundaries, and counting them using a Python script (total corpus: 16,250 different trigrams)
3. Normalised these counts with the number of `IP-MAT`-tokens in each place (= text length)
4. Removed trigrams that occur ≤ 5 times in the whole corpus ($\rightsquigarrow$ 2939 different trigrams)



EXPERIMENT

5 Computed the logarithmic transform to reduce the effect of the Zipfian distribution $\rightsquigarrow$ less strong effect of frequent trigrams that appear in all texts $\Rightarrow 42 \times 2939$ frequency matrix

	A	B	C	D	E	F	G	H	I
1	**trigram**	zzz zzz np-sbj	zzz np-sbj vfin	zzz zzz conj	pp zzz zzz	np-acc zzz zzz	zzz conj np-sbj	vfin np-acc zzz	np-sbj vfin np-acc
2	01_vpa	0,480938416	0,301319648	0,21260997	0,19354839	0,18255132	0,16202346	0,134897361	0,113636364
3	02_ctl	0,406685237	0,269266481	0,29897864	0,19220056	0,161559889	0,211699164	0,131847725	0,126276695
4	03_pft	0,425404945	0,245524297	0,27109974	0,2114237	0,158567775	0,198635976	0,119352089	0,111679454
5	04_aal	0,355775578	0,207920792	0,23366337	0,16369637	0,141914191	0,148514851	0,094389439	0,082508251
6	05_pal	0,397072278	0,238792315	0,21774931	0,18938701	0,163769442	0,136322049	0,115279048	0,082342177
7	06_clc	0,51028481	0,375	0,14794304	0,14240506	0,171677215	0,095727848	0,148734177	0,151107595
8	07_pst	0,385888502	0,25261324	0,25609756	0,19947735	0,193379791	0,18554007	0,140243902	0,141114983

EXPERIMENT

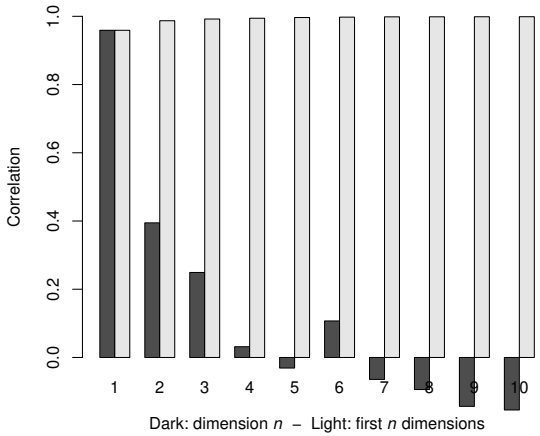
6 From this: 42 × 42 distance matrix (cosine similarity)

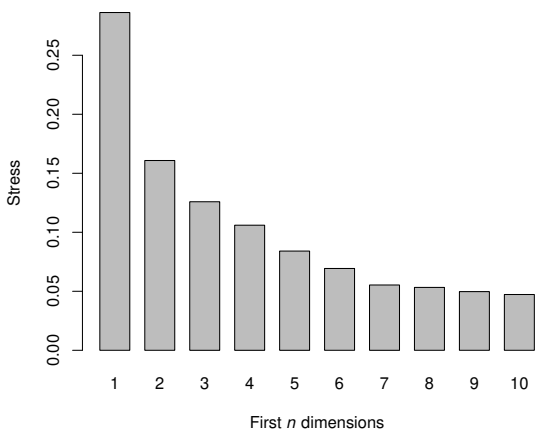
	A	B	C	D	E	F
1		01_vpa	02_ctl	03_pft	04_aal	05_pal
2	01_vpa	0	0.0037133423362639695	0.009187919779522158	0.007077090957890886	0.005165822923544616
3	02_ctl	0.0037133423362639695	0	0.0068853699892309495	0.005813541014137669	0.006317640990545215
4	03_pft	0.009187919779522158	0.0068853699892309495	0	0.00640064527394546	0.013919607350866503
5	04_aal	0.007077090957890886	0.005813541014137669	0.00640064527394546	0	0.011322497407409093
6	05_pal	0.005165822923544616	0.006317640990545215	0.013919607350866503	0.011322497407409093	0

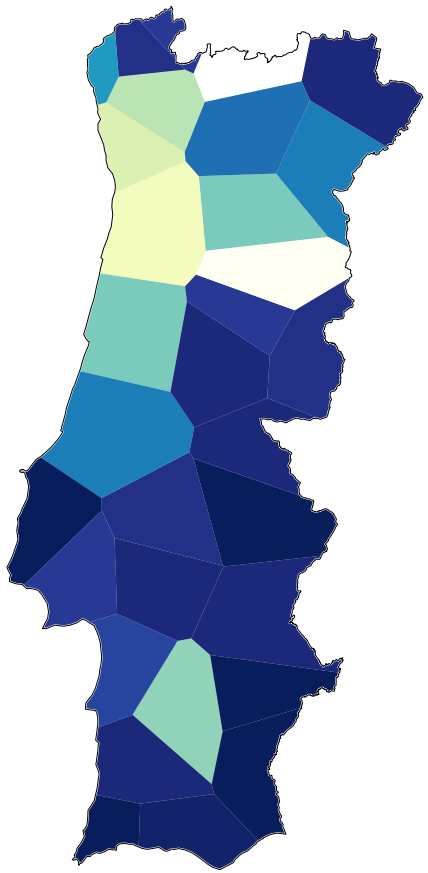
EXPERIMENT

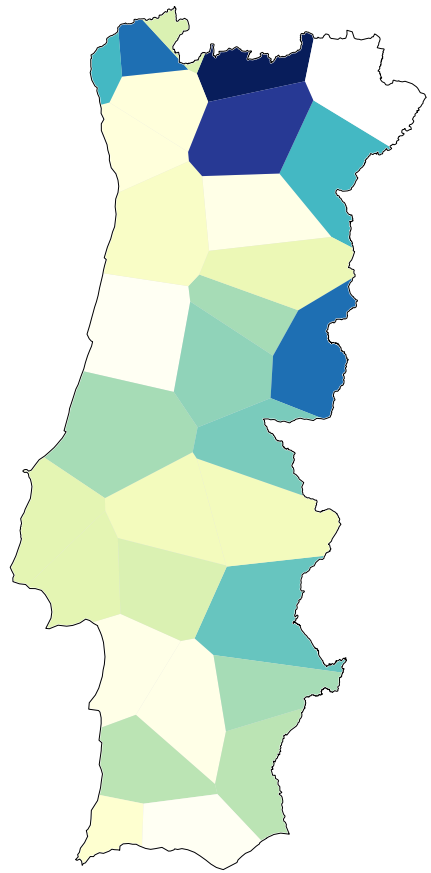
7 Multidimensional Scaling (using Gabmap, Nerbonne et al. 2011),

(first 2 dimensions: r = 0.99)





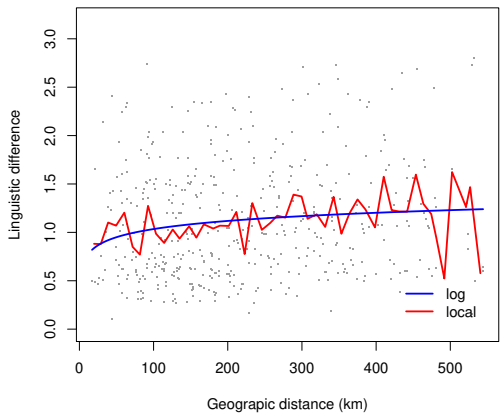


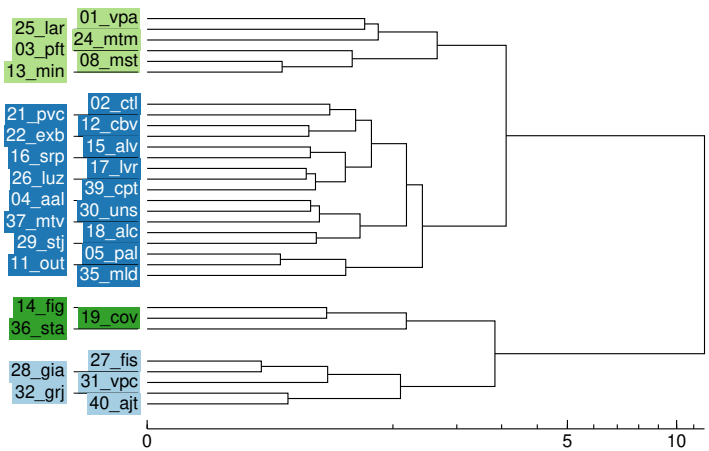


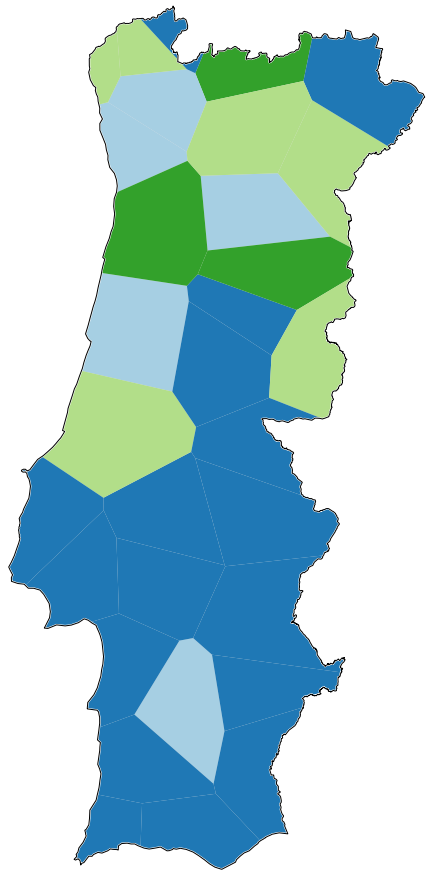
GHENT UNIVERSITY

EXPERIMENT

8 Clustering (Ward's method, 4 clusters, using Gabmap, Nerbonne et al. 2011)



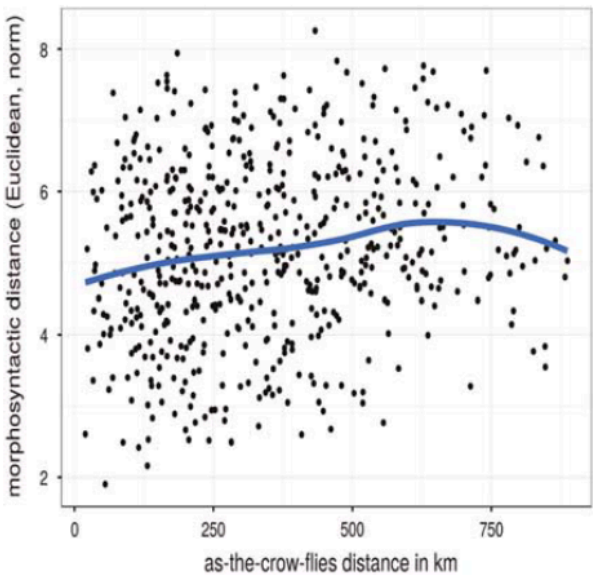




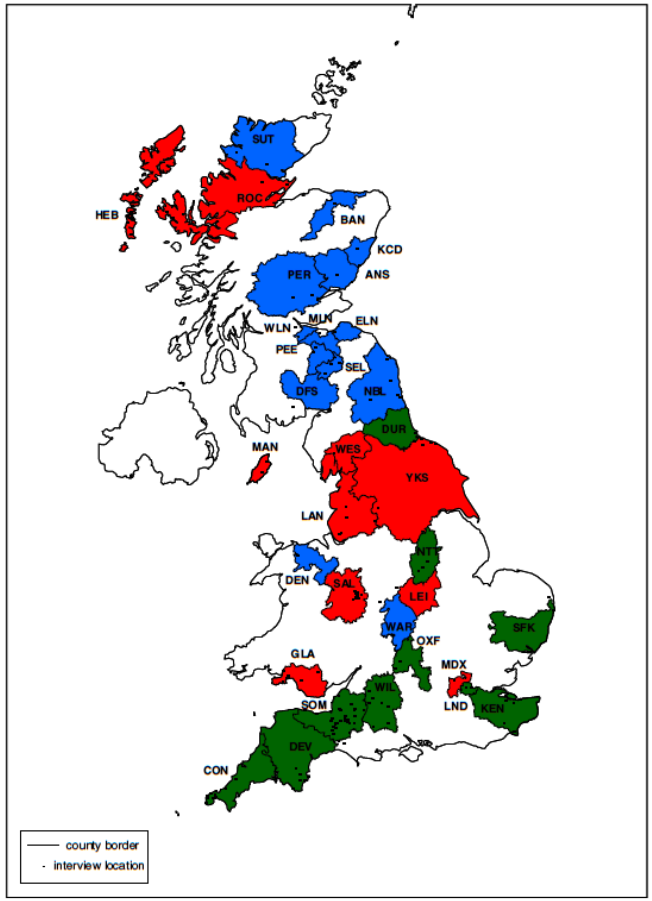
COMMON 'PROBLEM'

"In linguistic terms, this means that syntax is more prone to nonareal variation. Similar syntactic distributions are, to some extent, areally discontinuous."

(Birkenes/Fleischer 2021: 157)[1]



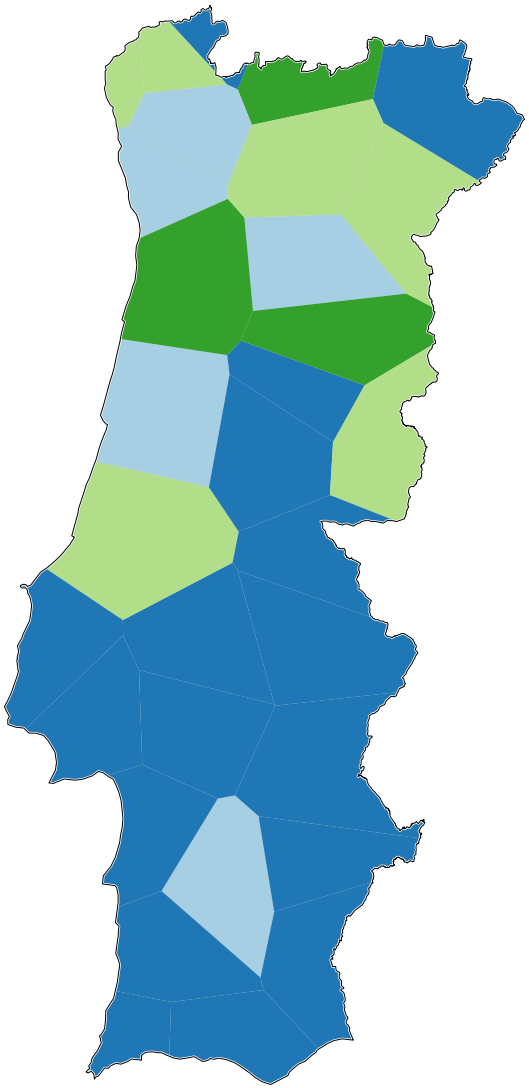
(Wolk/Szmrecsanyi 2016:9)

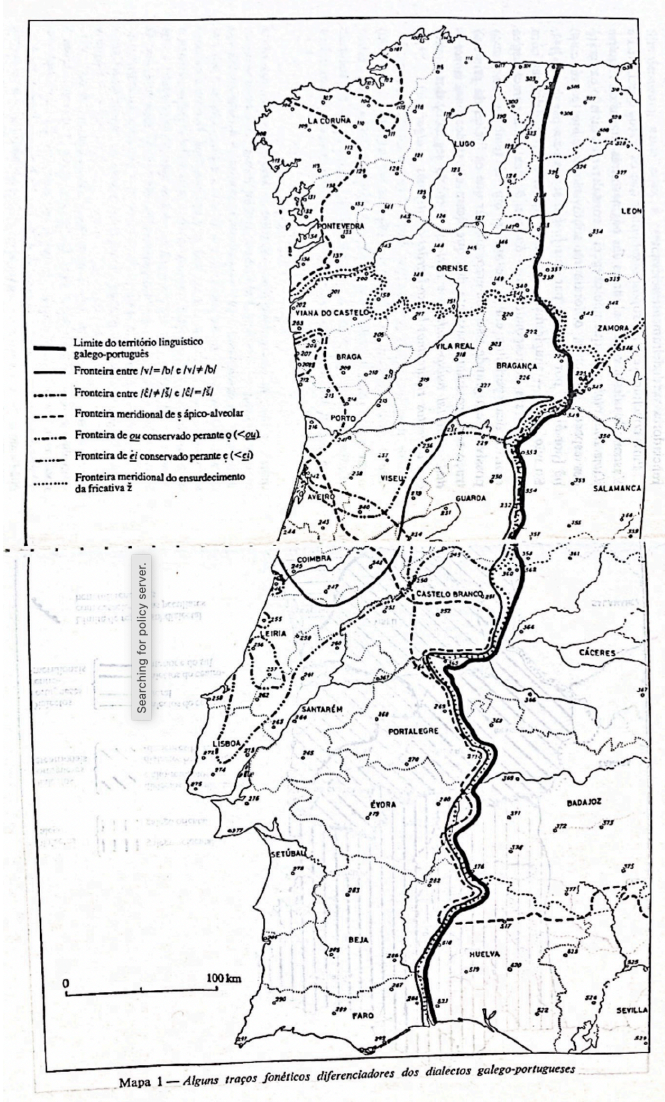


(Wolk/Szmrecsanyi 2016:8)

[1]=Comparison of syntactic variables in atlas data with character trigrams in Wenker-questionnaires (cf. also Birkenes 2020) to establish whether phonological and syntactic variation align, geospatially.

THOUGH...





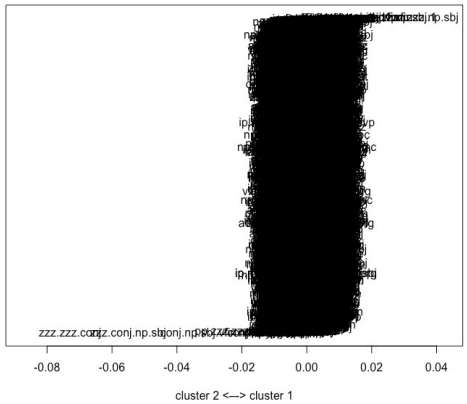
Mapa 1 — *Alguns traços fonéticos diferenciadores dos dialectos galego-portugueses*

"The principal dialect division in Portugal is north versus south, with an approximate transition to the north of Coimbra." (Lipski 2017: 503)

(Lindley Cintra 1983: 160–1)

FEATURES

Cluster 1:2



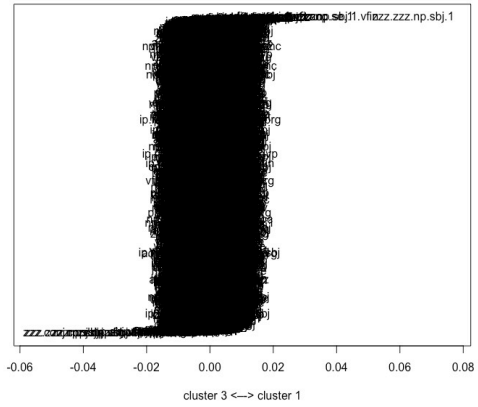
Most distinctive:

- zzz.zzz.np-sbj-1 >
- zzz.conj.np-sbj-1 >
- zzz.np-sbj-1.vfin >
- np-sbj-1.vfin.np-se-1 >
- vfin.np-se-1.np-acc >
- ip-smc.zzz.zzz >

Least distinctive:

- zzz.zzz.np-sbj <
- zzz.np-sbj.vfin <
- np-sbj.vfin.np-acc <
- vfin.np-acc.zzz <
- vfin.zzz.zzz

Cluster 1:3



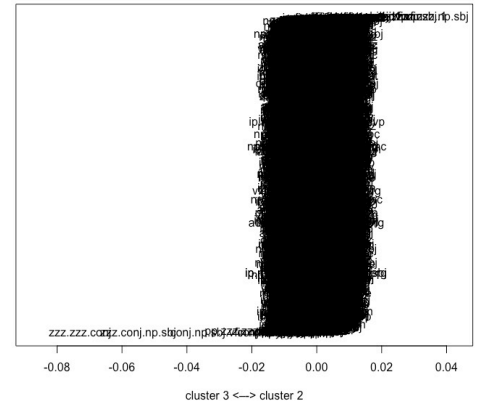
Most distinctive:

- zzz.zzz.np-sbj-1 >
- zzz.np-sbj-1.vfin>
- np-sbj-1.vfin.np-se-1>
- vfin.np-se-1.np-acc>
- vfin.np-1.zzz >
- zzz.conj.np-sbj-1

Least distinctive:

- zzz.zzz.np-sbj <
- zzz.conj.np-sbj <
- zzz.np-sbj.vfin <
- conj.np-sbj.vfin <
- zzz.zzz.conj

Cluster 2:3



Most distinctive:

- zzz.zzz.np-sbj >
- zzz.zzz.np-sbj-1 >
- zzz.np-sbj-1.vfin >
- zzz.np-sbj.vfin >
- zzz.np-sbj.advp >
- np-1.zzz.zzz

Least distinctive:

- zzz.zzz.conj <
- zzz.conj.np-sbj <
- conj.np-sbj.vfin <
- pp.zzz.zzz <
- conj.np-sbj.advp

GHENT UNIVERSITY

(REMAINING) CHALLENGES

- Syntactic variables: only positive data, no linking between possible syntactic alternatives

  $\longrightarrow$ Maybe fixable with first applying some CorpusSearch coding queries

  – bringing top-down back?

- Linking surface patterns to underlying parameters

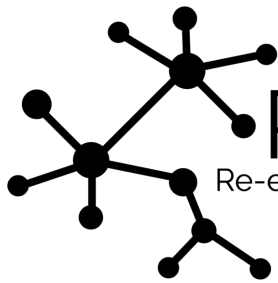
SUMMING UP

SUMMING UP

- Some syntactic phenomena resist elicitation
  $\rightsquigarrow$ traditional top-down methods of dialectology have their limits for syntax
- Alternative: where possible, complement questionnaire/atlas data with bottom-up data from spontaneous speech
- Room for improvement: need to develop better methods for extracting and using quantitative and qualitative information in those data

Thank you!

Research Foundation
Flanders
Opening new horizons



REEDS
Re-examining dialect syntax

REFERENCES

Barbiers, S., H.Bennis, G. De Vogelaer, M.Devos & M. van der Ham. 2005. *Syntactische Atlas van de Nederlandse Dialecten. Deel I.*. Amsterdam: AUP.

Birkenes, Magnus Breder. 2020. Zur Klassifikation der niederdeutschen Dialekte anhand von Buchstaben-*n*-Grammen. *Niederdeutsches Jahrbuch* 143: 86–113

Birkenes, Magnus Breder & Jürg Fleischer. 2020. Niederdeutsch in Hessen: Das Zeugnis syntaktischer Strukturen. *Niederdeutsches Jahrbuch* 143: 32–48.

Birkenes, Magnus Breder & Jürg Fleischer. 2021. Syntactic vs. phonological areas: A quantitative perspective on Hessian dialects. *Journal of Linguistic Geography* 9: 142–161

Blancquaert, E. 1948. Na meer dan 25 jaar dialect-onderzoek op het terrein. *Koninklijke Vlaamse Academie voor Taal- en Letterkunde* III. 5–62.

van Bree, C. 1996. H*istorische taalkunde*. Leuven/Amersfoort: Acco.

Farasyn, M. 2021. Een bijzondere constructie in het Frans-Vlaams: *'t Maakt nuus we gingen naar de frèreschool*. https://www.de-lage-landen.com/article/een-bijzondere-discoursmarkeerder-in-het-frans-vlaams-t-maakt-nuus-we-gingen-naar-de-frereschool

Lindley Cintra, L.F. 1983. *Estudos de Dialectologia Portuguesa*.. Lisboa: Sá de Costa Editora.

Lipski, J.M. 2018. Dialects of Spanish and Portuguese. Chpt. 30 of C.Boberg, J.Nerbonne & D.Watt (eds.), The Handbook of Dialectology, 498–509. Oxford: Wiley-Blackwell.

Martins, A.M. 2000–. CORDIAL-SIN: Corpus Dialectal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects. Lisboa: Centro de Linguística da Universidade de Lisboa. http://www.clul. ulisboa.pt/en/10-research/314-cordial-s

Magro, C. 2010. When CORDIAL Becomes Friendly: Endowing the CORDIAL Corpus with a Syntactic Annotation Layer. LREC 2010: 3705–3711.

Nerbonne, J. R.Colen, C.Gooskens, P.Kleiweg & T.Leinonen (2011). Gabmap — A Web Application for Dialectology. Dialectologia Special Issue II, 65-89.

Postma, G.J. 2002. De enkelvoudige clitische negatie in het Middelnederlands en de Jespersen cyclus. *Nederlandse Taalkunde* 7: 44–82.

Randall, Beth. 2005. CorpusSearch 2 User's Guide. University of Pennsylvania.

Séguy, J. (1971) 'La relation entre la distance spatiale et la distance lexicale', Revue de Linguistique Romane 35, 335–357.

Szmrecsanyi, B. 2011. Corpus-Based Dialectometry: A methodological sketch. *Corpora* 6(1): 45–76.

Szmrecsanyi, B./L.Anderwald. 2018. Corpus‐Based Approaches to Dialect Study. In: C.Boberg/J.Nerbonne/D.Watt (eds.), *The Handbook of Dialectology*, 300–313. Oxford: Blackwell.

Wolk, C. 2014. Integrating Aggregational and Probabilistic Approaches to Dialectology and Language Variation. Ph.D. dissertation, Albert-Ludwigs-Universität Freiburg i. Br.

Wolk, C./B.Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. In Marie-Hélène Côté, Remco Knooihuizen and John Nerbonne (eds.), *The future of dialects*, 225–244. Berlin: Language Science Press. DOI:10.17169/langsci.b81.152