Measuring distance-based coherence

Benedikt Szmrecsanyi (KU Leuven)

Word count: 6,301

Abstract

This paper utilizes Variation-Based Distance & Similarity Modeling (VADIS; see Szmrecsanyi, Grafmiller & Rosseel 2019) to quantify the distance and similarity between lects (in our case study, nine international varieties of English) as a function of the (non-)correspondence of the ways in which language users choose between different ways of saying the same thing. The method can be used to yield distance-based coherence (DBC) measurements. These measurements operationalize coherence between sub-datasets as being proportional to the extent to which these sub-datasets yield similar distance relations. The method is innovative because coherence is not measured on the level of "surfacy" variant rates but at the deeper level of how variation is conditioned probabilistically. Analysis of large corpus-derived variationist datasets about three syntactic variables in the grammar of English (the dative alternation, the genitive alternation, and the particle placement alternation) shows that coherence across alternations is surprisingly precarious, that coherence across the spoken-written distinction is surprisingly robust, and that coherence across lines of evidence (as customary in comparative sociolinguistics) is measurable, but not perfect.

Acknowledgments

A grant from the Research Foundation Flanders (FWO) is gratefully acknowledged (grant # G.0C59.13N). Thanks go to Matt Hunt Gardner and to two anonymous reviewers for helpful feedback on an earlier version of this manuscript.

1. Introduction

This paper quantifies the distance and similarity between lects (in our case study, nine international varieties of English) as a function of the (non-)correspondence of the ways in which language users choose between different ways of saying the same thing. We will be specifically concerned with grammatical variation, and the the notion of 'coherence' which takes center stage in this paper draws inspiration (a) from comparative sociolinguistics (e.g. Tagliamonte 2001), thanks to the types of empirical evidence we consider; (b) from the literature on experience-based/usage-based probabilistic grammar (e.g. Bresnan & Ford 2010), thanks to the gradient probabilistic differences in linguistic choice making that we will be interested in; and (c) from dialectometry (e.g. Szmrecsanyi 2013) and quantitative typology (e.g. Cysouw 2013), thanks to the method's reliance on calculating distances between varieties for the sake of gauging coherence.

What exactly does "coherence" mean the in the comparative variationist analysis of international varieties of English? We investigate three different aspects of distance-based coherence (henceforth: DBC):

1. $DBC_{alternation}$ is defined as being proportional to the extent to which the probabilistic conditioning of different grammatical alternations (i.e., grammatical variables) predicts similar linguistic distances between varieties.
2. DBC_{medium} is defined as being proportional to the extent to which measurements in spoken language materials and in written language materials predict similar linguistic distances between varieties.
3. $DBC_{evidence}$ is defined as being proportional to the extent to which different so-called lines of evidence in comparative sociolinguistics (constraint significance, effect directions, constraint ranking) predict similar linguistic distances between varieties.

So concretely we are asking the following questions:

1. If two varieties (say, British English and Canadian English) turn out to be similar when we investigate a particular alternation (e.g. the dative alternation), will those varieties also turn out to be similar when we probe another alternation (e.g. the particle placement alternation) ($DBC_{alternations}$)?
2. If two varieties turn out to be similar when we restrict attention to grammatical choice-making in spoken production, will those varieties also turn out to be similar when we restrict attention to choice-making in written language production (DBC_{medium})?
3. If two varieties turn out to be similar when we investigate a particular line of evidence (e.g. constraint significance), will those varieties also turn out to be similar when we probe another line of evidence (e.g. constraint ranking) ($DBC_{evidence}$)?

Determining distance-based coherence as sketched above is based on measuring linguistic distance between varieties. To calculate these distances, the study marshals Variation-Based Distance & Similarity Modeling (VADIS) (see Szmrecsanyi, Grafmiller & Rosseel 2019) to quantify the similarity between varieties and dialects as a function of the correspondence of the ways in which language users choose between different ways of saying the same thing. VADIS draws on concepts and ideas developed in Comparative Sociolinguistics (see e.g. Tagliamonte 2001; Tagliamonte 2012: 162–173; Tagliamonte, D'Arcy & Louro 2016).

The method is based on determining the probabilistic conditioning of one or more linguistic variables. In the case study reported in this paper, the following alternations in the grammar of English will be investigated:

(1) The genitive alternation (see, e.g., Heller 2018; Heller, Szmrecsanyi & Grafmiller 2017)

a. *the country's economic crisis* (ICE-SIN, s2b-001) (the *s*-genitive)

b. *the economic growth of the country* (ICE-IND w2a-031) (the *of*-genitive)

(2) The dative alternation (see, e.g., Röthlisberger 2018a; Röthlisberger, Grafmiller & Szmrecsanyi 2017)

a. *I'd given Heidi my T-Shirt* (ICE-GB s1b-066) (the ditransitive dative variant)

b. *I'd given the key to Helen* (ICE-CAN s1a-058) (the prepositional dative variant)

(3) The particle placement alternation (see, e.g., Grafmiller & Szmrecsanyi 2018)

a. *I have not picked the book up in 2 years* (GloWbE CA) (verb-object-particle order)

b. *My 6yo [sic] old actually picked up the book* (GloWbE NZ) (verb-particle-object order)

These alternations are investigated in corpus material covering nine international varieties of English (British English, Canadian English, Irish English, New Zealand English, Hong Kong English, Indian English, Jamaican English, Philippine English, and Singapore English).

This paper leaves aside examining any community-internal social stratification of the above alternations given that across previous studies none have shown strong social effects (e.g. Jankowski & Tagliamonte 2014; Röthlisberger & Tagliamonte 2020; Tagliamonte 2014). Rather, the aim is to investigate if the conditioning of variation coheres across alternations, across the spoken-written medium distinction, and across lines of evidence. The notion of the 'linguistic community' is relevant to this paper insofar as the data points that will allows us to take distance measurements are individual varieties of English, which constitute speech communities at a macro level. We will have nothing to say about coherence *within* linguistic communities.

Analysis shows that $DBC_{alternations}$ is surprisingly precarious, that DBC_{medium} is surprisingly substantial, and that $DBC_{evidence}$ is measurable but not perfect.

This paper is structured as follows: Section 2 presents the data sources. Section 3 explains the method. Results will be presented in Section 4. Section 5 offers a discussion and concluding remarks.

2. Data

We analyze the genitive alternation dataset investigated by Heller (2018), the dative alternation dataset investigated by Röthlisberger (2018a), and the particle placement dataset investigated by Grafmiller and Szmrecsanyi (2018) (see examples (1) to (3) above). The datasets were created by tapping into the International Corpus of English (ICE; http://ice-corpora.net/ice/index.html) (Greenbaum 1991; Greenbaum 1996) and the Corpus of Global Web-based English (GloWbE; https://www.english-corpora.org/glowbe/) (Davies & Fuchs 2015) to investigate syntactic variation in the following nine varieties of English:

- British English (henceforth: BrE)
- Canadian English (CanE)
- Irish English (IrE)
- New Zealand English (NZE)
- Jamaican English (JamE)
- Singapore English (SgE)
- Indian English (IndE)
- Hong Kong English (HKE)
- Philippine English (PhlE)

ICE is a set of parallel, balanced corpora representing language use across a wide range of (standard) national varieties. Each ICE component contains 500 texts of approximately 2,000 words each, sampled from 12 spoken (e.g. face-to-face conversations, broadcast news) and written genres/registers (e.g. student essays, press news reports). GloWbE covers data collected from 1.8 million English language websites—both blogs and general web pages—from 20 different countries (approximately 1.8 billion words in all).

We include both what Kachru (1985; 1992) has called 'Inner Circle' varieties of English (BrE, IrE, CanE, and NZE) and 'Outer Circle' varieties of English (JamE, SgE, IndE, HKE, and PhlE). This distinction roughly corresponds to MacArthur's (1998) distinction between English as a Native Language (ENL) varieties (communities "in which the language is spoken and handed down as the mother tongue of the majority of the population"; Schneider 2011: 30), and English as a Second Language (ESL) varieties (communities "in which English has been strongly rooted for historical reasons and assumes important internal functions (often alongside indigenous languages), e.g. in politics (sometimes as an official or co-official language), education, the media, business life, the legal system, etc."; Schneider 2011: 30). According to the literature (see Röthlisberger & Szmrecsanyi 2019 for discussion), this is a very important dialect-typological distinction in English linguistics because substrate and contact influences as well as SLA universals (see e.g. Klein & Perdue 1997) shape the grammars of Outer Circle varieties but not Inner Circle varieties.

In the corpus materials genitive, dative, and particle placement variants were identified. Only variants that could be paraphrased by the competing variant with no semantic change were included. For reasons of space, we cannot review the definitions of the variable contexts in detail here; the reader is referred to the discussions in Heller (2018), Röthlisberger (2018a), and Grafmiller and Szmrecsanyi (2018). In a second step, after all interchangeable variants were identified in the materials (dative alternation: N = 13,171; genitive alternation: N = 13,798; particle placement alternation: N = 11,454 – note that varieties of English are represented by approximately equal token numbers, so roughly between 1,000 and 2,000 tokens per variety per alternation), each observation was annotated, manually or automatically, for constraints on syntactic variation, such as constituent length/weight in all three alternations, constituent animacy in the genitive and dative alternation, and so on. Again, for reasons of space we cannot discuss the annotation procedure in detail; the reader is referred to Heller (2018), Röthlisberger (2018a), and Grafmiller and Szmrecsanyi (2018).

The datasets and the *R* code used to carry out the analysis are publicly available at https://osf.io/3gfqn/. The dative dataset is also available as Röthlisberger (2018b).

3. Method: VADIS

The following is a condensed version of the more detailed description in Szmrecsanyi et al. (2019). VADIS is designed to measure the (dis)similarity of grammars. "Grammar" is understood here as a set of probabilistic grammars (or "variable grammars") conditioning a set of $N \geq 1$ alternations or variable phenomena (or "variables"). A probabilistic grammar specifies the set of constraints (a.k.a. predictors or "conditioning factors" in variationist sociolinguistics parlance) regulating a given alternation.

VADIS builds on methods developed in comparative sociolinguistics (see e.g. Tagliamonte 2001; Tagliamonte 2012: 162–173; Tagliamonte, D'Arcy & Louro 2016), which is a sub-discipline in variationist sociolinguistics that evaluates the relatedness between varieties based on the similarity of the conditioning of variation in these varieties. Comparative sociolinguists rely on "the three lines of evidence" to determine relatedness, which the VADIS method re-interprets in the following way:[1]

1. Are the same constraints significant across varieties (constraint significance)?
2. Do the constraints have the same strength across varieties (constraint strength)?
3. Is the constraint hierarchy similar (constraint ranking)?

VADIS draws inspiration from this literature and adapts the Comparative Sociolinguistics method so that it can be applied to datasets sampling (a) more than two varieties, and (b) more than one variable phenomenon at a time. This is accomplished through additional quantification. Practically speaking, VADIS consists of the following steps:

[1] In classical Comparative Sociolinguistics, the 2nd line of evidence and the 3rd line of evidence are conceptualized and measured differently than in VADIS: in Comparative Sociolinguistics, the 2nd line refers to the relative strength of factors, while the 3rd line of evidence refers to the ranking of factors within factor groups.

Step 1: Define, per alternation, the *p* most important constraints on variation. For the present study, we set $p = 8$ and so include the eight most important[2] predictors (across all varieties) for each alternation. The predictor sets thus generated are reported in Szmrecsanyi et al. (2019). Here we summarize as follows. Genitive alternation: possessor animacy, possessor length in words, possessum length in words, possessor NP expression type, final sibilancy in possessor, priming, semantic relation, possessor head frequency; dative alternation: weight ratio between recipient and theme, recipient pronominality, theme complexity, theme head frequency, theme pronominality, theme definiteness, recipient givenness, recipient head frequency; particle placement alternation: length of the direct object in words, definiteness of the direct object, givenness of the direct object, concreteness of the direct object, thematicity of the direct object, presence of a directional modifier, semantics, information-theoretic surprisal of the particle given the verb.

Step 2: Fit a series of mixed-effects logistic regression models, one per variety and alternation. The response variable is variant choice (e.g. *s*-genitive versus *of*-genitive), and the independent variables are the predictor sets identified in step 1.

Step 3: Based on the variety-specific regression models, determine cross-variety distance based on predictor significance. In this step, we define the probabilistic distance between two varieties as being proportional to the extent to which the varieties do *not* overlap with regard to which constraints significantly regulate variant choice (in the case study at hand, we set alpha = 0.05) . To exemplify, consider two hypothetical varieties A and B and five constraints a-e which regulate some variation phenomenon:

	variety A	variety B
constraint a	significant	significant

[2] See Szmrecsanyi et al. (2019) for how importance was determined.

constraint b	significant	not significant
constraint c	not significant	significant
constraint d	not significant	not significant
constraint e	significant	significant

Variety A and B agree on the significance of three constraints (a, d, e), and disagree with regard to two constraints. The distance between the two varieties is thus two out of five squared Euclidean distance points. Scaling this to an interval between 0 (no disagreement whatsoever) and 1 (maximal disagreement) yields, in the example at hand, a distance value of $2/5 = 0.4$.

Step 4: Based on the variety-specific regression models, determine cross-variety distance based on the magnitude of effects. To define the distance between the varieties, this step compares the extent to which the effect sizes of the constraints in the various regression models are dissimilar. This is done by calculating a distance matrix based on the model estimates (using Euclidean distance), regardless if they are significant. See Szmrecsanyi et al. (2019) for details.

Step 5: Fit a series of conditional random forest models, one per variety and alternation. To independently estimate the relative importance of the constraints, we use permutation-based variable importance rankings derived from conditional random forests (CRFs; Strobl, Malley & Tutz 2009). The response variable and independent variables in the models are the same as for the regression models in Step 2.

Step 6: Based on the variety-specific conditional random forest models, determine cross-variety distance based on the importance rankings of the predictors (3rd line of evidence). In this last step, we measure the probabilistic distance between two varieties simply as the inverse Spearman rank correlation between those varieties' respective variable importance rankings.

For example, consider the three hypothetical varieties A, B, and C with the constraint rankings below:

	rank variety A	rank variety B	rank variety C
constraint a	1	1	2
constraint b	2	3	4
constraint c	3	2	3
constraint d	4	4	1
constraint e	5	5	5

Varieties A and B show the greatest degree of similarity, with a correlation of $\rho = .9$ (inverse distance value: 1 - .9 = .1), while varieties A and C are least similar, with a correlation of $\rho = .3$ (inverse distance value: 1 - .3 = .7). Variety B is slightly more similar to variety C than variety A is ($\rho = .4$; inverse distance value: 1 - .4 = .6).

An R package (under development) which performs all the above calculations is available at https://github.com/jasongraf1/VADIS.

Distance values as calculated above can be arranged in so-called distance matrices. Distance matrices are the customary input in classical dialectometry (Séguy 1971; Goebl 1982; Nerbonne, Heeringa & Kleiweg 1999; Szmrecsanyi 2013) and function like distance tables in road atlases, which specify geographic distances between locations. For each alternation in our case study and for each of the three lines of evidence, we create one distance matrix. As we are exploring $n = 9$ varieties of English, this yields $n \times (n\text{-}1)/2 = 9 \times 8/2 = 36$ unique variety pairings.

```
      BrE    CanE    HKE    IndE    IrE    JamE    NZE    PhlE
```

CanE	0.000							
HKE	0.310	0.310						
IndE	0.548	0.548	0.238					
IrE	0.286	0.286	0.048	0.167				
JamE	0.095	0.095	0.262	0.452	0.262			
NZE	0.095	0.095	0.190	0.476	0.167	0.048		
PhlE	0.286	0.286	0.452	0.571	0.333	0.405	0.310	
SgE	0.214	0.214	0.310	0.429	0.167	0.286	0.167	0.095

Figure 1. VADIS distance matrix for the 3rd line of evidence (constraint ranking) in the particle placement alternation (all data included, 8 constraints considered). Scores range between 0 (maximal similarity) and 1 (maximal distance).

Figure 1 shows the distance matrix for the 3rd line of evidence (constraint ranking) in the particle placement alternation. All distances are scaled between 0 (no distance) and 1 (maximal distance). Consider now e.g. the pairing between BrE and NZE, which is associated with a comparatively small distance value of 0.095. In other words, BrE and NZE are very similar in terms of the constraint ranking in the particle placement alternation. By contrast, the distance between BrE and IndE is 0.548, which is considerably larger.

Distance matrices are informative but somewhat hard to process visually. There are, however, a number of techniques to visualize distance matrices. One of these is Multidimensional Scaling (MDS) (see e.g. Kruskal & Wish 1978), which reduces a higher-dimensional distance matrix to a lower-dimensional representation. MDS is therefore a dimension-reduction technique (originally developed in psychometrics) which translates distances between objects (in this case, linguistic distances between varieties of English) in high-dimensional space (note that in Figure 1, each variety is characterized by distances to 8 other varieties) into a lower-dimensional representation that can be visually depicted in two-dimensional plots in which distances between the lects are represented as proportionally as possible to the linguistic distances in the original high-dimensional distance matrix. In other words, proximity in an MDS plot indicates linguistic similarity.

Per alternation, we are initially dealing with three separate distance matrices (one per line of evidence), which could in principle be plotted separately. However, let's abstract away for the moment from individual lines of evidence by fusing[3] the three line-specific distance matrices, thus arriving at line-merged but alternation-specific distance matrices.

Figure 2 displays the corresponding MDS plots. Note that the axes (and the scaling that these depict) of MDS plots are typically not strongly interpreted, which is another way of saying that the absolute position of the data points in the plots does not matter much. What matters is proximity between data points: are two varieties that are comparatively close in e.g. the genitive alternation plot also close in e.g. the dative alternation plot? Cursory inspection of the plots reveal substantial differences between alternations (we revisit this issue in the next section), but also similarities – for instance, across all three alternations, IndE and PhlE are distant from the other varieties.

[3] Fusion is accomplished using the the fuse() function in R package analogue (see https://cran.r-project.org/web/packages/analogue/analogue.pdf). This creates a compromise matrix specifying mean distances.

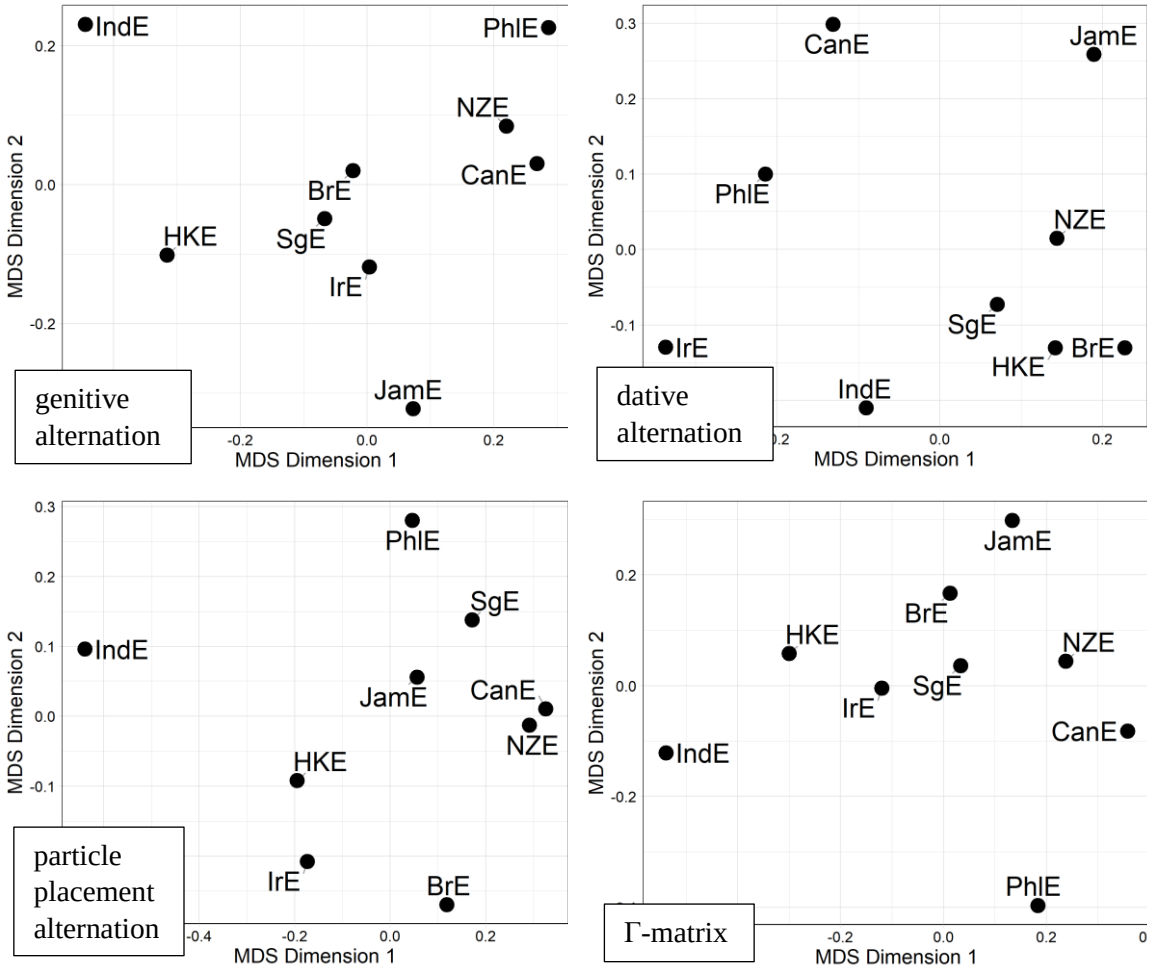

Figure 2. MDS representations of distance matrices. Upper left: compromise distances according to the genitive alternation. Upper right: compromise distances according to the dative alternation. Lower left: compromise distances according to the particle placement alternation. Lower right: MDS representation of the Γ-matrix (a single compromise distance matrix merged across all lines and alternations). Distances between data points in plot is proportional to probabilistic grammar distances between varieties.

We may now take a further aggregation step for the sake of raising the analysis of distance relationships to an even higher level of generalization. This we can accomplish by fusing the three alternation-specific-distance matrices into a single compromise distance matrix merged across all lines and alternations, or Γ-MATRIX for short. An MDS visualization of this Γ-matrix for the present study is to be found in the lower right plot in Figure 2. In that plot, all Inner Circle varieties are clustered in the top right-hand quadrant (IrE marginally so), with SgE – which according to the literature is an Outer Circle variety in the process of becoming an Inner Circle variety (Leimgruber 2013: 122) – forming part of that cluster. IndE and PhlE are outliers.

4. Results

With basic knowledge about how VADIS works under our belt, we may now take distance-based coherence measurements as follows.

4.1. Coherence across alternations: $DBC_{alternation}$

In this section, we are asking the following question: if, according to alternation A, two varieties are close in terms of how people choose between different ways of saying the same thing, will the two varieties also be close when the analysis is based on alternations B or C? In more visual terms, consider again Figure 1: what is the extent to which distance relationships in the individual MDS plots for the three alternations are similar to each other? This similarity is what $DBC_{alternations}$ measures.

overlap genitive alternation/dative alternation	$r = 0.05$ ($p = 0.41$)
overlap genitive alternation/particle alternation	**$r = 0.52$ ($p = 0.01$)**
overlap dative alternation/particle alternation	$r = 0.11$ ($p = 0.31$)

Table 1. $DBC_{alternation}$ measurements: Mantel correlation coefficients between fused distance matrices (combining all lines of evidence and based on all available data). Significant coefficients are bolded.

We specifically measure DBC by quantifying the overlap between alternation-specific distance matrices using the Mantel test (Levshina 2015: 348–349), which, based on permutation[4], yields correlation coefficients that range between 0 (no overlap) and 1 (total overlap). Note here that we do not correlate or compare the MDS plots – these are just visual aids. What we correlate via the Mantel test is the underlying high-dimensional distance matrices (which also means that the scaling of the axes in the MDS plot and similar issues do not play any role in our calculations).

Table 1 reports the resulting three correlation coefficients. In short, we find significant and substantial overlap between the genitive alternation and the particle placement alternation, while the dative alternation does not overlap with either one of the other alternations.

Going back to the MDS plots in Figure 1, which visually depicts the MDS plots subject to comparison here, one pattern that emerges is that the genitive and particle placement plots reveal a cluster of Inner Circle varieties (BrE, IrE, CanE, and NZE – located in the right half of the genitive alternation plot, and in the bottom right quadrant of the particle placement alternation plot). This cluster is absent from the dative alternation plot, which is one reason why the dative alternation shows little overlap with the other alternations.

4.2. Coherence across the spoken-written distinction: DBC_{medium}

The corpora we tap into to study syntactic variability cover both spoken language (e.g. face-to-face conversation, unscripted speeches) and written language (e.g. press news reports, web-based materials).

[4] Among other things, permutation fixes problems arising from the statistical dependence of elements within each of the matrices to be compared.

So the question that we are asking in this section is the following: If two varieties turn out to be close when we restrict attention to grammatical choice-making in spoken production, will those varieties also turn out to be close when we restrict attention to choice-making in written language production?

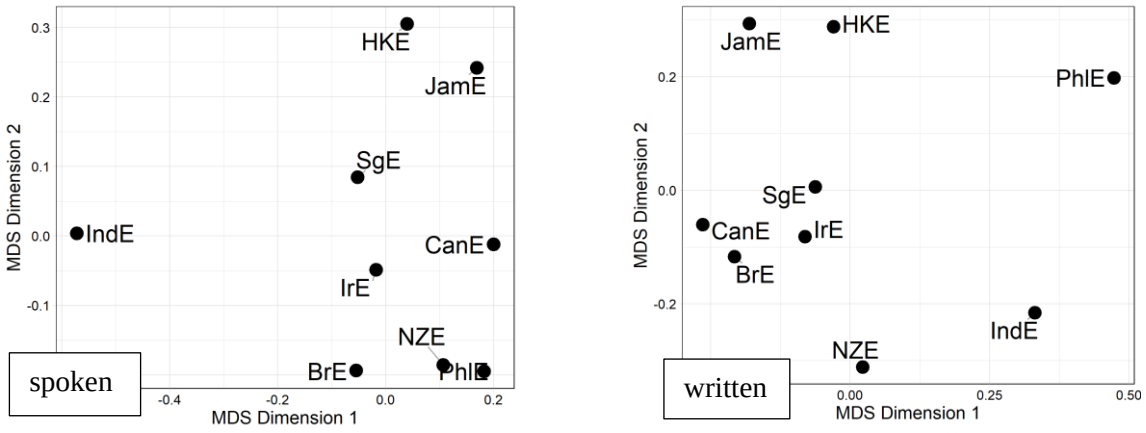

Figure 3. MDS representations of distance matrices. Left: Γ-matrix, spoken materials only. Right: Γ-matrix, written materials only. Distances between data points in plot is proportional to probabilistic grammar distances between varieties.

To address this question, we split up the datasets into spoken sub-datasets and written sub-datasets, and subsequently conduct separate VADIS analyses, which generate two different Γ-matrices (one for spoken production and the other one for written production). These matrices are plotted in Figure 3 via MDS. Visual inspection suggests some coherence, in that for example in both plots we find a reasonably tight cluster of Inner Circle varieties (BrE, IrE, CanE, and NZE).

To quantify DMC_{medium}, we turn to the Mantel test for the sake of measuring the correlation between the two distance matrices: $r = 0.56$ ($p = 0.02$). According to customary thresholds in interpreting correlation coefficients in the social sciences, we thus see "substantial to very strong" (De Vaus 2002:

272) overlap between the distance matrices. In other words, there is a good deal of coherence between spoken and written language production.

4.3. Coherence across lines of evidence: $DBC_{evidence}$

Recall that the VADIS method, which we use in this paper to assess coherence, draws on concepts and ideas originally developed in Comparative Sociolinguistics (see e.g. Tagliamonte 2001; Tagliamonte 2012: 162–173; Tagliamonte, D'Arcy & Louro 2016). One of these ideas is that variational similarity between lects may be evaluated by considering three lines of evidence:

1. Are the same constraints significant across varieties?
2. Do the constraints have the same strength across varieties?
3. Is the constraint hierarchy similar?

The $DBC_{evidence}$ measurements in this section quantify the extent to which the three lines of evidence really measure the same thing or not: If two varieties turn out to be close if we restrict attention to a particular line of evidence (e.g. constraint significance), will those varieties also turn out to be close when we probe another line of evidence (e.g. constraint strength)?

	genitive alternation	dative alternation	particle alternation
overlap 1st line/2nd line	***r* = 0.41 (*p* = 0.03)**	*r* = 0.12 (*p* = 0.34)	***r* = 0.36 (*p* = 0.05)**
overlap 1st line/3rd line	*r* = 0.07 (*p* = 0.36)	*r* = -0.01 (*p* = 0.50)	*r* = 0.25 (*p* = 0.13)
overlap 2nd line/3rd line	***r* = 0.47 (*p* = 0.03)**	*r* = -0.15 (*p* = 0.77)	***r* = 0.68 (*p* = 0.00)**

Table 2. Mantel correlation coefficients between line-of-evidence-specific distance matrices. Significant coefficients are bolded.

Rather than creating compromise distance matrices per alternation (which merge line-specific distance matrices), let us consider in this section three separate distance matrices per alternation, one for each line of evidence. We then calculate Mantel correlation coefficients between these line-specific distance matrices. The results are shown in Table 2. We note, first, that the dative alternation is the odd one out once again in that none of the lines overlap with each other in this alternation. Second, the genitive alternation and the particle placement alternation are similar in that they both show moderate but significant overlap between the first line of evidence (constraint significance) and the second line of evidence (constraint strength), as well as substantial overlap between the second line of evidence and the third line of evidence (constraint ranking). We do not see significant overlap anywhere between the first line of evidence (constraint significance) and the third line of evidence (constraint ranking). In short, $DBC_{evidence}$ scores suggest that there is some – but not perfect – overlap (or redundancy) in the three lines of evidence.

5. Discussion and Conclusion

This paper has utilized Variation-Based Distance & Similarity Modeling (VADIS) to quantify the distance and similarity between lects (in our case study, nine international varieties of English) as a function of the (non-)correspondence of the ways in which language users choose between different ways of saying the same thing. In the case study at hand, we specifically looked at the English gentive, dative, and particle placement alternations. As we have seen, VADIS is inspired by concepts and ideas developed in comparative sociolinguistics, by work in experience-based/usage-based probabilistic grammar, and by techniques widely used in dialectometry and quantitative typology.

Crucially, the method can be used to take so-called distance-based coherence (DBC) measurements. We have investigated three different varieties of DBC, guided by three different research questions: if two

varieties (say, British English and Canadian English) turn out to be similar when we investigate a particular alternation (e.g. the dative alternation), will those varieties also turn out to be similar when we probe another alternation (e.g. the particle placement alternation) ($DBC_{alternation}$)? If two varieties turn out to be similar when we restrict attention to grammatical choice-making in spoken production, will those varieties also turn out to be similar when we restrict attention to choice-making in written language production (DBC_{medium})? And finally: if two varieties turn out to be similar when we investigate a particular line of evidence (e.g. constraint significance), will those varieties also turn out to be similar when we probe another line of evidence (e.g. constraint ranking) ($DBC_{evidence}$)?

As to coherence between alternations ($DBC_{alternations}$), we saw that the distance matrices derived from the genitive and particle placement alternations do overlap substantially, but the dative alternation distance matrix does not overlap with any of the other distance matrices. In short, the dative alternation is an outlier. The deeper theoretical question here is whether grammar (or the variable parts of grammar) is essentially a collection of independent and/or independently conditioned alternations, or whether alternations actually "agree", as it were, about differences between varieties: if, say, BrE and SgE are close according to genitive alternation measurements, will they also be close according to dative alternation measurements? And so on. Our analysis suggest that we are dealing with a mixed picture: we see both coherence and non-coherence. With regard to non-coherence, it is unexpected that and unclear why the dative alternation does not pattern with the other alternations – all three alternations under study are, after all, constituent order alternations that are constrained by similar factors (constituent length, animacy, and so on). Further work is needed to investigate why the dative alternation is different from the other alternations. In any event, one is reminded here of Guy (2013), who investigates if people consistently use stigmatized or prestige variants, and finds that it is surprisingly hard to demonstrate correlations in the behavior of variables, even if they are generally thought to vary along the same social dimensions. Likewise, our finding that alternations do not necessarily cohere perfectly calls into question conceptions of grammar that consider grammar the orderly aggregation of binary alternations.

As to coherence across the spoken-written distinction (DBC_{medium}), the data show that there is substantial overlap between distance matrices calculated on the basis of spoken materials only and distance matrices calculated on the basis of written materials only. Thus, if two varieties turn out to be similar when we restrict attention to spoken production, those varieties will also tend to be similar when we restrict attention to choice-making in written language production. This sort of coherence is maybe not entirely unexpected to those who have always believed that spoken and written language production go underlyingly by the same rules. But substantial written-spoken overlap will perhaps be a bit surprising to all those who believe that (vernacular) speech does or should have a special status, because It is in this style where variation is thought to be most systematic: it is "the style in which the minimum attention is given to the monitoring of speech" (Labov 1972: 208). Our results suggest that spoken language is maybe not that special – but then again, of course, to be fair in the Labovian perspective difference between spoken and written would not be expected in the operation of constraints but rather in the frequency of variants. Others might have supposed that spoken language varieties should be different because written language varieties are subject to stylistic and prescriptive norms in a way that spoken language varieties are not (see, e.g., Hinrichs, Szmrecsanyi & Bohmann 2015). Lastly, substantial written-spoken coherence defies a plausible research hypothesis that spoken language varieties should be different because the production of spoken syntax is subject to processing and production constraints and biases (e.g Hawkins 1994; MacDonald 2013) in a way that the production of written syntax is not (though we hasten to add that $r = 0.56$ does leave some room for differences based on processing/production constraints).

As to coherence across lines of evidence ($DBC_{evidence}$), analysis shows that there tends to be overlap between the 1st line of evidence (constraint significance) and the 2nd line of evidence (effect size), as well as between the 2nd line of evidence and the 3rd line of evidence (constraint ranking). This is true for the genitive alternation and the particle placement alternation; the distance matrices generated on the basis of data from the dative alternation do not overlap at all (this confirms the outlier status of the dative alternation established in our analysis of $DBC_{alternations}$). All this underlines the wisdom of designing the

Comparative Sociolinguistics method such that it taps into three lines of evidence that show partial overlap but do not measure *exactly* the same thing.

In summary, we saw that $DBC_{alternations}$ is surprisingly precarious, that DBC_{medium} is surprisingly substantial, and that $DBC_{evidence}$ is measurable but not perfect.

Note now that one inherent limitation of the case study reported here is that the datasets we analyzed do not provide information about social factors such as speaker/writer gender, speaker/writer age, and so on. This prevented us from investigating a fourth type of distance-based coherence, which we may label DBC_{social}: If two varieties turn out to be similar when we restrict attention to materials produced by female speakers, will those varieties also turn out to be similar when we restrict attention to materials produced by male speakers? If two varieties turn out to be similar when we restrict attention to materials produced by younger speakers, will those varieties also turn out to be similar when we restrict attention to materials produced by older speakers? And so on. Exploring these questions is an exciting avenue for future research.

References

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213. https://doi.org/10/cb3tn2.

Cysouw, Michael. 2013. Disentangling geography from genealogy. In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in Language and Linguistics*. Berlin, Boston: DE GRUYTER. http://www.degruyter.com/view/books/9783110312027/9783110312027.21/9783110312027.21.xml (31 January, 2015).

Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1). 1–28. https://doi.org/10/gdh9t3.

De Vaus, D. A. 2002. *Analyzing social science data*. London ; Thousand Oaks, Calif: SAGE.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.

Grafmiller, Jason & Benedikt Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *Language Variation and Change* 30(03). 385–412. https://doi.org/10/gf4p2w.

Greenbaum, Sidney. 1991. ICE: the International Corpus of English. *English Today* 7(04). 3. https://doi.org/10/d4762p.

Greenbaum, Sidney. 1996. *Comparing English worldwide: the International Corpus of English*. Oxford, New York: Clarendon.
Guy, Gregory R. 2013. The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics* 52. 63–71. https://doi.org/10.1016/j.pragma.2012.12.019.
Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge; New York: Cambridge University Press.
Heller, Benedikt. 2018. *Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation Across Varieties of English*. Leuven: KU Leuven PhD dissertation.
Heller, Benedikt, Benedikt Szmrecsanyi & Jason Grafmiller. 2017. Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation Across Varieties of English. *Journal of English Linguistics* 45(1). 3–27. https://doi.org/10/gf7nv8.
Hinrichs, Lars, Benedikt Szmrecsanyi & Axel Bohmann. 2015. Which-hunting and the Standard English relative clause. *Language* 91(4). 806–836. https://doi.org/10/gf7nqs.
Jankowski, Bridget L. & Sali A. Tagliamonte. 2014. On the genitive's trail: data and method from a sociolinguistic perspective. *English Language and Linguistics* 18(02). 305–329. https://doi.org/10/gf7nsm.
Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: the English language in the outer circle. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the World: Teaching and Learning the Language and Literatures*, 11–30. Cambridge: Cambridge University Press.
Kachru, Braj B. (ed.). 1992. *The Other tongue: English across cultures* (English in the Global Context). 2nd ed. Urbana: University of Illinois Press.
Klein, Wolfgang & Clive Perdue. 1997. The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research* 13. 301–347. https://doi.org/10/d5t5c4.
Kruskal, Joseph B. & Myron Wish. 1978. *Multidimensional Scaling*. Newbury Park, London, New Delhi: Sage Publications.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia press.
Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139225755. http://ebooks.cambridge.org/ref/id/CBO9781139225755 (13 September, 2018).
Levshina, Natalia. 2015. *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4. 1–16. https://doi.org/10/gbfpt3.
McArthur, Tom. 1998. *The English Languages*. Cambridge: Cambridge University Press.
Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit Distance and Dialect Proximity. In David Sankoff & Joseph Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, v–xv. Stanford: CSLI Press.
Röthlisberger, Melanie. 2018a. *Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation*. Leuven: KU Leuven PhD dissertation. https://lirias.kuleuven.be/handle/123456789/602938.
Röthlisberger, Melanie. 2018b. *The dative dataset of World Englishes*. KU Leuven. https://doi.org/10.5281/zenodo.2553357.
Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4). 673–710. https://doi.org/10/gddnmm.
Röthlisberger, Melanie & Benedikt Szmrecsanyi. 2019. Dialect Typology: Recent Advances. In Stanley D Brunn & Roland Kehrein (eds.), *Handbook of the Changing World Language Map*, 1–26. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73400-2_133-1. http://link.springer.com/10.1007/978-3-319-73400-2_133-1 (11 December, 2019).

Röthlisberger, Melanie & Sali A. Tagliamonte. 2020. The social embedding of a syntactic alternation: Variable particle placement in Ontario English. *Language Variation and Change* 32(3). 317–348. https://doi.org/10.1017/S0954394520000174.

Schneider, Edgar. 2011. *English Around the World: an Introduction*. [S.l.]: Cambridge University Press.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35. 335–357.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348. https://doi.org/10/b3hg89.

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: a study in corpus-based dialectometry* (Studies in English Language). Cambridge, [England] ; New York: Cambridge University Press.

Szmrecsanyi, Benedikt, Jason Grafmiller & Laura Rosseel. 2019. Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes. *Frontiers in Artificial Intelligence* 2. 23. https://doi.org/10/ggcgp7.

Tagliamonte, Sali. 2001. Comparative sociolinguistics. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *Handbook of Language Variation and Change*, 729–763. Malden and Oxford: Blackwell.

Tagliamonte, Sali. 2012. *Variationist sociolinguistics change, observation, interpretation*. Malden, Mass.: Wiley-Blackwell. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=819316.

Tagliamonte, Sali A. 2014. A comparative sociolinguistic analysis of the dative alternation. In Rena Torres-Cacoullos, Nathalie Dion & André Lapierre (eds.), *Linguistic variation: Confronting fact and theory*, 297–318. London, New York: Routledge.

Tagliamonte, Sali A., Alexandra D'Arcy & Celeste Rodríguez Louro. 2016. Outliers, impact, and rationalization in linguistic change. *Language* 92(4). 824–849. https://doi.org/10/gdg6vt.