Morphosyntactic variation in Dutch dialects: theoretical vs. geographical distance

Jeroen van Craenenbroeck[1,2] & Marjo van Koppen[2,3]
[1]KU Leuven/CRISSP, [2]Meertens Institute, [3]Utrecht University/UiL-OTS

October 11, 2021

Abstract

In this paper we explore a methodology for comparing different analyses of a set of morphosyntactic variation data from Dutch dialects. The analyses we focus on are a parametric account on the one hand and a purely geography-based one on the other. Based on an experiment that uses the k-nearest neighbor classification we are able to precisely and explicitly weigh and compare the two accounts, and by projecting the outcome of the experiment back onto a geographical map, we gain further insight into the relative strengths and weaknesses of the approaches under consideration.

1 Introduction

The past twenty years have seen the birth of a whole host of syntactically oriented dialect atlas projects across Europe and the US. As a result, we now have detailed empirical overviews of morphosyntactic variation in dialects of, to name but a few, Dutch (Barbiers et al. 2005, 2008), Scandinavian (Lindstad et al. 2009), Swiss German (Glaser and Bart 2015, Buchelli and Glaser 2002), Alemannic (Brandner 2015), Hessian (Fleischer et al. 2015), and North American English (Zanuttini et al. 2018). The projects just mentioned differ from one another in a large number of respects: the precise linguistic phenomena that were focused on, the number and sociolinguistic profile of the informants, the methodology used to elicit the data, etc. However, there is one constant that came out of each and every project, and that is the sheer amount of variation that was uncovered. Traditionally, dialectologists focused on lexical, phonological, and—to a lesser extent—morphological variation; the syntactic properties of closely related dialect varieties were typically assumed to be invariant. The data that have emerged from the microcomparative approach have shown this assumption to be misguided, as there is an abundance of such morphosyntactic variation.

This influx of new data raises fundamental theoretical questions. Generative linguists typically assume that the (morphosyntactic) variation found in natural language is not arbitrary and unlimited, but systematic and subject to grammatically determined principles. Such an approach works well when confronted with a limited number of phenomena in a limited number of languages, but when faced with variation involving hundreds of variables in hundreds of dialects, it meets substantial methodological challenges. Within the dialectometric tradition, an approach to language variation in which computational and statistical methods are applied in dialectological research, on the other hand, geographical distance is assumed to play a key role in accounting for the variation patterns. An example of this line of thinking is Nerbonne and Kleiweg (2007)'s Fundamental Dialectological Postulate, which states that geographically proximate varieties tend to be more similar (linguistically) than distant ones. In other words, two nearby language varieties share certain linguistic features in the same way that the speakers of those varieties share cultural events or customs.

In this paper we want to contribute to this debate by introducing and comparing two different ways of looking at the same data set of morphosyntactic variation data: a formal-parametric analysis and a purely geography-based one. Although we will draw an explicit comparison between the two approaches, our goal in this paper is not to declare The Ultimate Winner between the two. Rather, we want to examine to what extent and in what aspects these approaches overlap or are complementary, and what kinds of tools we can use to establish this. The paper is organized as follows. The next section introduces the data that form the empirical core of the paper. We focus on ten different dialect phenomena from Dutch dialects (Barbiers et al. 2006). Section 3 introduces the two analyses, and in section 4, we introduce a machine learning method for comparing them. Section 5 looks at ways in which we can visualize our results, while section 6 sums up and concludes.

2 The data

The data set that forms the empirical basis for this paper is the one introduced by Van Craenenbroeck and van Koppen (2021). It concerns a set of ten dialect phenomena found in a subset of the dialects of Dutch spoken in Belgium, the Netherlands, and the northern tip of France. The first is the phenomenon known as complementizer agreement (van Koppen 2017), whereby a complementizer introducing a finite embedded clause can agree in person and/or number with the subject of that clause. An example from the dialect of Gistel is given in (1), where the complementizer *on* agrees in number with the third person plural subject *Bart en Lieske*.

(1) **O-n** Bart en Lieske in t paradijs levn
if-pl Bart and Lieske in the paradise live
'If Bart and Lieske are living in paradise, ...' Gistel, Barbiers et al. (2006)

The second phenomenon is clitic doubling, whereby a pronominal subject—a strong pronoun in particular—can be doubled by a clitic pronoun, which cliticizes onto either the complementizer (in embedded clauses) or the finite verb (in inverted main clauses) (see e.g. Haegeman (1992); van Craenenbroeck and van Koppen (2002); de Vogelaer (2005). An example is given in (2).

(2) da-**ze** **zaaile** lachen.
that-they$_{clitic}$ they$_{strong}$ laugh
'that they are laughing.' Wambeek

The third phenomenon, dubbed short *do* replies by Van Craenenbroeck (2010), involves short contradictory answers that contain the verb *doen* 'to do'. In the Berlare example in (3) the B-speaker contradicts A's negative statement by means of an affirmative *do*-reply.

(3) A: IJ zal nie komen. B: **IJ doet**.
he will not come he does
'A: He won't come. B: Yes, he will.' Berlare, Barbiers et al. (2006)

The fourth construction under investigation here is well-known from negative concord languages, i.e. the use of a negative clitic in addition to the main negator to express a single semantic negation (see Haegeman and Breitbarth (2014)). An example from the dialect of Tielt is given in (4), where the negative clitic *en* is combined with the negative adverb *nie* 'not'.

(4) K **en** goa nie noar schole.
I neg go not to school
'I'm not going to school.' Tielt, Barbiers et al. (2006)

The fifth phenomenon is the addition of subject clitics to the polarity markers *yes* and *no* (see Van Craenenbroeck (2010)). In the Malderen example in (5), B's short affirmative answer to A's yes/no-question contains not only the affirmative polarity marker *ja* 'yes', but also a subject clitic that matches the person and number of the intended full clausal reply.

(5) A: Wilde nog koffie, Jan? B: **Ja-k**.
want.you part coffee Jan Yes-I
'A: Do you want some more coffee, Jan? B: Yes.' Malderen, Barbiers et al. (2006)

The sixth phenomenon concerns the morphology or etymology of the *there*-expletive in Dutch dialects. Specifically, while most dialects use a locative form like the standard language, some resort to what appears to be a phonologically reduced form of the third person singular neuter pronoun *het* 'it' (see Van Craenenbroeck (2019, 2020)). An example from the West Flemish dialect of Brugge is given in (6).

(6) **T** en goa niemand nie dansn.
it neg goes no.one not dance
'There will be no one dancing.' Brugge, Barbiers et al. (2006)

The seventh phenomenon we look at is the form of the comparative complementizer. In particular, certain dialects (such as that of Oostkerke illustrated in (7)), use the coordinating disjunction *of* 'or' to introduce the standard of comparison.

(7) Zie peist daj eer ga thuis zijn **of** ik.
she thinks that.you sooner go home be or I
'She thinks you'll be home sooner than me.' Oostkerke, Barbiers et al. (2006)

The eighth construction on our list has to do with the possibility of eliding the locative expletive pronoun *er* 'there' in embedded clauses and inverted main clauses (see Bennis (1986:214), Zwart (1992), Klockmann et al. (2015)). While many dialects allow for such a deletion, there are some that do not, as illustrated in (8).

(8) dat ***(er)** in de fabrieke nen jongen werkte
that there in the factory a boy worked
'that a boy worked in the factory Lapscheure, Haegeman (1986:3)

Ninthly, some dialects allow for the combination of a definite determiner and a demonstrative in contexts of NP-ellipsis (Corver and van Koppen 2018) as in (9), while Standard Dutch disallows this type of demonstrative doubling.

(9) **De die** zou k ik wiln op eetn.
the those would I_{clitic} I_{strong} want up eat
'I would like to eat those.' Merelbeke, Barbiers et al. (2006)

Finally, some Dutch dialects display a phenomenon reminiscent of so-called quirky verb second in Afrikaans (de Vos 2006), whereby two verbs seem to have raised to clause-initial position instead of the usual one. As shown in (10), in Dutch dialects this phenomenon affects imperative clauses, whereby the main verb appears as a (finite) imperative in second position, preceded by an infinitival form of the aspectual auxiliary *go* or *come*.

(10) **Gon haalt** die bestelling ne keer!
go_{inf} get_{imp} that order a time
'Go get that order!' (Ghent)

The goal of this brief overview was not to examine any of these constructions in any detail—we refer to the references mentioned above for in-depth descriptions and theoretical analyses—but rather to give the reader a general impression of the type of variation data we are concerned with in this paper. What makes this set interesting in light of the questions raised in the previous section is their geographical distribution. As shown in Figure 1, all ten of these phenomena are concentrated in the south west of the language area, but at the same time there are large differences between their individual distributions.[1] The overlap in distribution makes it not inconceivable that there is a shared grammatical or geographical principle underlying their presence or absence in a particular location, but at the same time the large disparaties in distribution might lead one to question such an account and hypothesize that the correlations

[1]The following abbreviations are used in Figure 1: CA = complementizer agreement, CD = clitic doubling, SDR = short *do* replies, NEG = negative clitic, CYN = clitics on *yes* and *no*, EXPL-T = the use of *it* as an expletive, COMPR = the use of *of* 'or' as a comparative marker, ER-OBL = no *there*-deletion in inversion and embedded clauses, THE-THAT = determiner-demonstrative doubling, GO-GET = quirky V2-like imperatives.

are spurious. In the next section we examine two possible analyses in more detail.

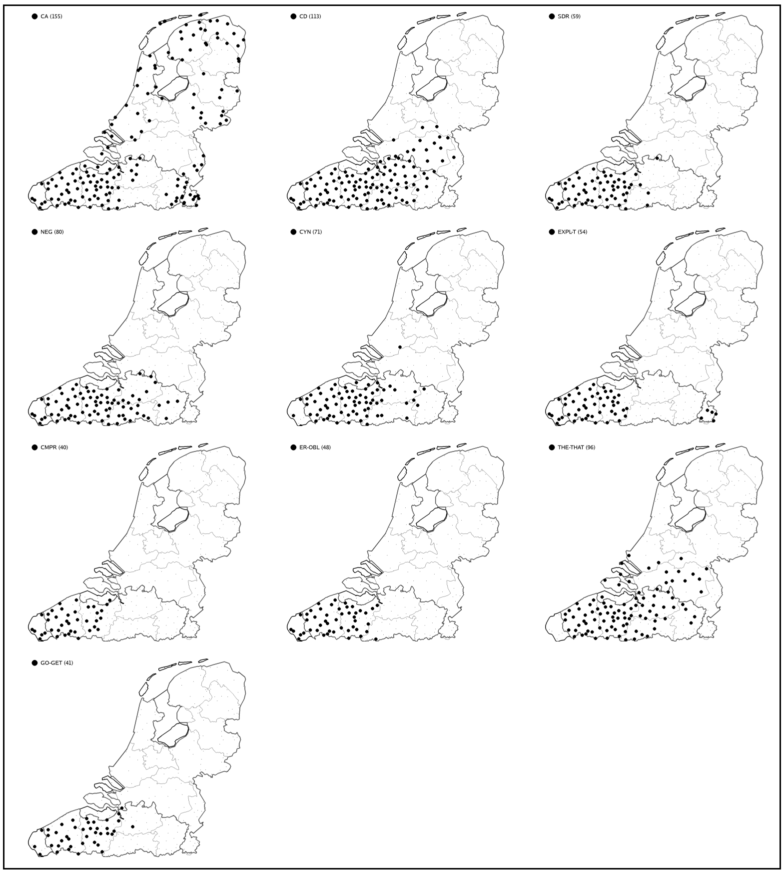

Figure 1: Geographical distribution of the ten dialect phenomena

3 Two possible analyses

As indicated in section 1, this paper wants to compare two types of analyses of the data introduced in section 2: a formal-theoretical one that tries to reduce the variation to a limited number of more abstract linguistic parameters, and a geography-based one that uses geographical distance as the determining factor in accounting for the variation. In the next two subsections we introduce these analyses in more detail.

3.1 The parametric analysis

The specific formal-linguistic analysis we will use as a point of comparison in this paper is the one developed by Van Craenenbroeck and van Koppen (2021) (henceforth VCVK). Discussing this analysis in detail and providing evidence in support of it would lead us too far afield, and so in this section we focus on introducing its main ingredients, and in particular those aspects of the analysis that are claimed to be responsible for the attested variation. For a fully worked-out account of this analysis, we refer the reader to the original paper.

VCVK present a parametric account of the data introduced in the previous section that is built on a prior quantitative-statistical analysis. In particular, they first apply a Correspondence Analysis to the raw data shown in Figure 1. Correspondence Analysis is a principal component method that can be applied to tables containing categorical data (for general discussion, see Greenacre 2007 and Levshina 2015:chapter 19). It reduces a data set containing many, possibly correlated variables to a smaller number of uncorrelated dimensions. Based on this dimension reduction, VCVK propose that the morphosyntactic variation shown in Figure 1 is the result of the interaction between three linguistic parameters. The first is what they call the AgrC-parameter:

(11) **AgrC-parameter:**
C {does/does not} have unvalued ϕ-features.

This parameter specifically regulates the occurrence or non-occurrence of complementizer agreement in a dialect. VCVK follow Van Koppen (2017) and Haegeman and van Koppen (2012) in assuming that complementizer agreement is the overt reflex of unvalued ϕ-features on C undergoing Agree with the subject. As a result, dialects that have such a ϕ-Probe on C display complementizer agreement, and dialects that do not lack complementizer agreement.

Note that this first parameter sets one phenomenon, complementizer agreement, apart from all the others. In other words, VCVK hypothesize that the distribution of complementizer agreement is unconnected and hence orthogonal to that of the remaining nine phenomena introduced in section 2. The second parameter, however, does link up two constructions. It is formulated in (12).

(12) **D-parameter:**
DP {does/does not} have a split left periphery.

The idea here is that certain dialects have an extended left periphery in their nominal domain, while others do not. This additional structural space can serve as a landing site for certain movement operations, thus making it possible for certain constructions to arise that would be disallowed in dialects that lack a split left periphery in the DP. Two such constructions are clitic doubling and determiner-demonstrative doubling. With respect to the first, VCVK adopt a so-called big DP-analysis of pronominal doubling, whereby the doubler and the doublee start out as one nominal constituent (see also Belletti (2005), Uriagereka (1995), Laenzlinger (1998), Grohmann (2000), Poletto (2008), Kayne (2005)). Specifically, following Van Craenenbroeck and van Koppen (2008), they argue that the clitic is the result of DP-internal movement of ϕP into the extended left periphery of the DP. As should be clear from the above discussion, such a movement operation can only occur in dialects that have a positive setting for the parameter in (12). Interestingly, Barbiers et al. (2016) argue on independent grounds for a highly similar analysis of determiner-demonstrative doubling: what looks like a determiner is actually the spell-out of a ϕP that has raised into the extended left periphery of the DP. As a result, VCVK assume these two phenomena to be regulated by a single parameter, the one in (12). The null hypothesis, then, would be that the distribution of these two phenomena would be identical: if the D-parameter is set to 'yes', both clitic doubling and determiner-demonstrative doubling should occur, while if it is set to 'no', neither should occur. As even a cursory comparison between maps 2 and 9 in Figure 1 shows, however, this null hypothesis is not met. This raises the question of how to determine the setting of the D-parameter: does it suffice for one of the two phenomena to be present—and if so, does it matter which one?—or should they both be attested for the parameter to be set to 'yes'? VCVK argue, partly based on a possible confound in the question methodology used to elicit determiner-demonstrative doubling, and partly based on historical data, that clitic doubling should be seen as the key indicator: a dialect has a positive setting for the D-parameter if and only if it has clitic doubling.

The third and final parameter that VCVK propose bundles together the remaining seven empirical phenomena listed in section 2. It mirrors the D-parameter, but applies at the clausal level:

(13) **C-parameter**
CP {does/does not} have a split left periphery.

The reasoning here parallels the one outlined above with respect to the D-parameter: dialects with a positive setting for the C-parameter provide more structural space in their clausal left periphery, and so constructions that specifically target or make use of this extra structural space are sensitive to the setting of this parameter. Once again, we have to make our discussion of the formal analysis of the seven phenomena that fall under this parameter schematic and brief, and we refer to VCVK and the references mentioned there for a more detailed and worked out account. With respect to the negative clitic, short *do*-replies, and the occurrence of clitics on 'yes' and 'no', Van Craenenbroeck (2010) argues that they all spell out or make use of a high left-peripheral polarity phrase (PolP), and the existence of such a projection requires a positive setting for the parameter in (13). The use of *t* 'it' as an expletive and the obligatory vs. optional occurrence of the locative expletive in inversion and embedded clauses is argued to be a reflex of a split versus an unsplit CP-domain by Van Craenenbroeck (2020): he argues that the *t*-expletive is not a pronoun, but rather a main clause complementizer spelling out the Force-head, while the optional deletion of the locative expletive is the reflex of another locative expression moving into the canonical subject position (Klockmann et al. 2015), an option which is only allowed if the left periphery is rich enough to host such a movement operation. That leaves the quirky V2-like *come/go get*-constructions and the use of *of* 'or' as a comparative marker. With respect to the former, VCVK argue that *come* and *go* have been grammaticalized and spell out a functional head higher than the one hosting the finite (imperative) verb. The use of *of* 'or' as a standard marker on the other hand is indicative of a lack of syncretism between the conjunction introducing conditional clauses and the one introducing comparative clauses, suggesting that in dialects that use *of* 'or' as a standard marker there are two separate C-layers for conditionals and comparatives, while in the syncretic dialects those features are bundled on a single C-head. In short, all of the seven remaining phenomena can be meaningfully linked to the presence or absence of structural space in the clausal left periphery, i.e. to the C-parameter as defined in (13). Just as was the case with the D-parameter, though, the non-identical geographical distribution of these seven constructions (see Figure 1) raises the question of how to determine the setting of this parameter in a specific dialect. VCVK argue that polarity plays a crucial role in informing the language-learning child of the correct setting of this parameter. More specifically, the C-parameter is set to 'yes' in a particular dialect if and only if at least one of the following polarity-related phenomena is attested in that dialect: the negative clitic, short *do*-replies, or clitics on 'yes' and 'no'.

This concludes our overview of VCVK's parametric account of the morphosyntactic variation introduced in section 2. It reduces said variation to the interplay between three parameters: the AgrC-parameter, the D-parameter, and the C-parameter. The first one specifically targets complementizer agreement, the second one bundles together clitic doubling and determiner-demonstrative doubling, while the third one covers the remaining seven phenomena. Together, these three binary parameters project eight different dialect groups. Membership of these groups is determined by the presence or absence of complementizer agreement (AgrC-parameter), the presence or absence of clitic doubling (D-parameter), and the presence or absence of at least one of the three polarity-related constructions (C-parameter).

3.2 The geographical analysis

The second analysis we want to use in the comparison in the next section is in a sense a much simpler one. It starts from the same intuition that also underlies Nerbonne and Kleiweg (2007)'s Fundamental Dialectological Postulate, which states that geographically proximate varieties tend to be more similar (linguistically) than distant ones. In other words, it assumes that geographical distance is a central component to understanding linguistic—in our case: morphosyntactic—variation, and so we expect linguistic features/constructions/phenomena/etc. to cluster geographically. Whether or not two linguistic varieties are alike does not depend on their grammatical system—i.e. their parameter settings, as per the previous analysis—but on their geographical location, in particular their vicinity to one another. Neighboring dialects share certain linguistic features in the same way that the speakers of those dialects share certain customs or cultural traditions. One way of interpreting this more generally would be to take geographical proximity as a proxy for the degree of language contact: speakers of nearby dialects are more likely to enter into contact with one another, thus increasing the chances that they will adopt characteristic features of one another's speech.

As should be clear from looking at the maps in Figure 1, the data under investigation in this paper seem quite amenable to such a geographical approach: with the exception of complementizer agreement, which has a much wider distribution, all phenomena are clustered in the southwestern part of the language area. Taking geographical distance to be a determining factor in understanding this variation thus seems to be a highly plausible hypothesis.

This concludes our introduction of the two analyses of the data introduced in section 2. We now proceed in the next section to an explicit comparison of the two approaches.

4 Comparing the two accounts

In this section we set out to compare the two analyses introduced in the previous section, by gauging their predictive power. We do so by means of an experiment with k-nearest neighbor (kNN) classification (Daelemans and Van den Bosch 2005).[2] This method tries to predict values of a variable based on other, memorized instances of that variable. More concretely, suppose we want to predict whether or not dialect X displays linguistic phenomenon Y. The kNN-method would then look at one or more other dialects that are similar to dialect X—i.e. its nearest neigbors—, determine whether or not they have phenomenon Y, and copy the majority outcome to dialect X. What is interesting from the point of view of the current paper is that the measure used to determine these nearest neighbors can vary. For example, dialects can be similar to one another because they are geographically local, but also because they share certain parameter settings, or even because of a combination of the two. This makes the kNN-method ideally suited for our present purposes: it can determine how well the analyses sketched in the previous section can predict the variation data introduced in section 2.

The experiment we have set up is of the type 'leave one out', whereby each of the 267 dialects in turn acts as the unknown, to-be-predicted value, while the remaining 266 dialects serve as the known, memorized instances from which the nearest neighbors can be drawn. These experiments are carried out for each of the ten dialect phenomena introduced in section 2, and in each case in three experimental runs: (1) using only geographical location—i.e. longitude and latitude—as the predictive feature, (2) using the binary values of the AgrC-parameter, the D-parameter, and the C-parameter, and (3) using a combination of the two. The outcome of (one run of) one experiment is a set of 267 classifications into $+$ ('the phenomenon occurs in this dialect') or $-$ ('the phenomenon does not occur in this dialect'). One way of summarizing such data would be to divide the number of correct classifications—i.e. the sum of the true positives and the true negatives—by the total number of classifications, i.e. 267, but this is too crude a measure in cases where the occurrence of a phenomenon is rare. For instance, the use of *of* 'or' as a standard marker only occurs in 40 of the 267 dialects. A method that always predicts this phenomenon to be absent would still get 227 of the 267 cases correct, an accuracy of 85%, in spite of the fact that none of the positive cases were correctly identified. In order to counter this effect, we computed the Area Under the ROC Curve (AUC) of the positive class (Fawcett 2004). This measure takes into account both the true positive rate—the number of true positives divided by the total number of cases—and the false positive rate—the number of false positives divided by the total number of classifications. It yields a value between 0 and 1, whereby 0.5 or lower indicates chance behavior and 1 perfect predictions. For more technical and mathematical details regarding the AUC-measure, we refer the reader to Fawcett (2004).

With this much as background, we can now turn to the results of our experiments. Table 1 lists the AUC value for each of the ten dialect phenomena, and each of the three experimental runs: the one where the nearest neighbors are selected based on geographical location, the one where we use parameter settings, and the third option, where we use both.[3]

When discussing the results, we will set aside those pertaining to complementizer agreement (CA) and clitic doubling (CD), i.e. the greyed-out rows in Table 1. Recall from subsection 3.1 that we used the distribution of these two phenomena to set the values of the AgrC- and the D-parameter respectively. It should come as no surprise, then, that a classification based on parameter values makes perfect predictions with respect to these two phenomena, while a location-based one fares worse (especially in the case of complementizer agreement, which has a distribution that is much less geographically homogeneous than the other phenomena, see Figure 1). This means our discussion for the remainder of the paper will be based on the remaining eight phenomena, none of which serves as the direct input for setting a

[2] All calculations were carried out in R (R Core Team 2014) using the class-package (Venables and Ripley 2002).

[3] Recall that we use the following abbreviations: CA = complementizer agreement, CD = clitic doubling, SDR = short *do* replies, NEG = negative clitic, CYN = clitics on *yes* and *no*, EXPL-T = the use of *it* as an expletive, COMPR = the use of *of* 'or' as a comparative marker, ER-OBL = no *there*-deletion in inversion and embedded clauses, THE-THAT = determiner-demonstrative doubling, GO-GET = quirky V2-like imperatives.

	location	parameters	location$+$parameters
CA	0.750	1.000	1.000
CD	0.929	1.000	0.997
SDR	0.895	0.940	0.900
CYN	0.906	0.909	0.918
NEG	0.912	0.955	0.931
CMPR	0.880	0.943	0.880
EXPL-T	0.928	0.914	0.930
GO-GET	0.840	0.916	0.816
THE+THAT	0.818	0.850	0.848
ER-OBL	0.860	0.959	0.871
AVERAGE	0.880	0.923	0.887
SD	0.038	0.035	0.041

Table 1: AUC-values of the kNN-experiment

parameter value.[4] Overall, it seems clear that a classification based on the grammatical parameters of Van Craenenbroeck and van Koppen (2021) fares better than a geography-based one: in six out of the eight cases, the parameters-only option yields the highest AUC-value (while the location-only one never does), and this holds even for phenomena that were not taken into account at all when determining the setting of the grammatical parameters, like the use of *of* 'or' as standard marker (CMPR) or the obligatory nature of the *there*-expletive (ER-OBL).[5] This general intuition is confirmed by the results of a two-tailed paired t-test, a summary of which is given in Table 2.

		t-value	p-value
location	parameters	-2.365	0.033
parameters	location$+$parameters	1.898	0.079
location	location$+$parameters	-0.353	0.729

Table 2: Results of a two-tailed paired t-test for the three experimental runs

The test reveals that a classification based on the parameters proposed by Van Craenenbroeck and van Koppen (2021) is significantly different from one that only uses geographical parameters (longitude and latitude), and that neither account differs significantly from one that uses a combination of both types of information. At the same time, we should not be too quick to dismiss the predictive power of the location-based account altogether. One aspect of the experiment we have not focused on so far is the value of the hyperparameter k. The value of k determines how many neighbors are taken into consideration in determining the missing value in the experiment. For example, if $k = 1$ in a location-only run of the experiment, the algorithm looks at the one dialect that is geographically closest to the one we are trying to classify, and it copies over the value of that dialect. If $k = 3$, it looks at three such dialects and copies over the majority value of these three, etc. (And in case of a tie—which is only possible when k is even—it randomly chooses one of the two values.) The parameter values of Van Craenenbroeck and van Koppen (2021) differ from the geographical info—i.e. longitude and latitude—in that they are not numeric, but categorical. For the kNN-algorithm, this means that there are only two possible distance measures

[4]Note that the numbers for average AUC-value and standard deviation in the bottom two rows of Table 1 also do not take the first two rows into account.

[5]Recall from subsection 3.1 that the C-parameter is set based on the values of the three polarity-related phenomena only.

between two dialects: either they have the exact same parameter setting, or they have a different one. This means that when trying to determine the nearest neighbor of a dialect with a certain parameter setting—say [$+AgrC$, $-D$, $+C$]—there is a tie between all dialects that have that same parameter setting, and all of them are taken into consideration, regardless of the value of k. In other words, when k is set to 1 (as it has been so far), the location-only run of the experiment considers only one other dialect in trying to determine the missing value—in particular the dialect that is geographically closest to the one with the missing value—while the parameters-only run takes into consideration all dialects with the same parameter setting. While in the latter case this seems justified—we consider the groups defined by these parameter settings to be natural classes, and it makes little sense to try and identify one dialect as being more representative of that class than the others—one might expect the predictive power of the location-only run of the experiment to go up if it is allowed to consider more than one nearest neighbor. In order to test this hypothesis we reran the experiment, but with values for k ranging from 1 tot 10.[6] Table 3 summarizes the results of those experiments.

		k	t-**value**	p-**value**
location	parameters	1	-2.365	0.033
		2	-2.351	0.036
		3	-1.350	0.206
		4	-1.566	0.142
		5	-1.094	0.267
		6	-1.145	0.273
		7	-1.079	0.299
		8	-1.224	0.242
		9	-1.224	0.242
		10	-1.231	0.240

Table 3: Results of a two-tailed paired t-test with different values for k

While the parameters-only account continues to outperform the location-only account in terms of the raw AUC-numbers, this difference is no longer statistically significant for values of $k = 3$ and upwards. In other words, when given enough neighboring dialects to consider, a geography-based account can increase its predictive power up to a level comparable to that of the parametric account.[7] This might lead one to hypothesize that both accounts meet a certain basic threshold in terms of their predictive power—i.e. that both provide a reasonable account for the variation data introduced in section 2—and that maybe there is a degree of complementarity between the two, that they account for different parts or aspects of the data set. In order to test this double hypothesis, we first set out to create an analysis that we could use as baseline. Given the substantial skew in terms of frequency in nine of the ten dialect phenomena—only complementizer agreement is fairly evenly distributed across the 267 dialects—we used a frequency-based classification: a run of the experiment whereby each missing value was filled in by simply copying the most frequent value for that variable in the entire data set.[8] The results of this experiment are shown in Table 4.

Note that in this table we are not listing AUC-values for the three experimental runs, but percentages of correct predictions (i.e. the sum of true positives and true negatives divided by the total number of

[6] For completeness' sake, we also ran a version of the experiment with $k = 1$ whereby the algorithm randomly chose one dialect (with the relevant parameter setting) in the parameters-only run of the experiment. In this scenario, the parametric account continued to outperform the geography-based one, with one glaring exception: the AUC-value for the quirky V2-like imperatives (GO-GET) dropped to 0.407, i.e. below chance level. While we believe this is an artefact of the specific dialect that was chosen by the algorithm, it is something we want to explore further in future research.

[7] Unsurprisingly, there is a limit to the number of dialects one wants to consider in the geography-based account: when too many dialects are taken into consideration—say, 50 or more—the predictive power rapidly decreases again, and the difference with the parametric account becomes significant once more.

[8] We implemented this baseline scenario as a kNN-experiment with location as the predictive feature and with k set to 266.

	location	parameters	baseline
CA	76.03%	100.00%	58.05%
CD	92.88%	100.00%	57.68%
SDR	92.13%	94.38%	77.90%
CYN	92.13%	93.26%	73.41%
NEG	92.13%	95.13%	70.04%
CMPR	93.63%	90.26%	85.02%
EXPL-T	95.13%	91.76%	79.78%
GO-GET	89.89%	89.14%	84.64%
THE+THAT	82.77%	85.02%	64.04%
ER-OBL	91.76%	93.26%	82.02%
AVERAGE	91.20%	91.53%	77.11%

Table 4: Percentages of correct predictions made by the two analyses vs. a frequency-based baseline

classifications). Given that the AUC-measure is specifically designed to counter the effect of frequency, it heavily penalizes our baseline account and yields AUC-values of 0.5, i.e. chance level, throughout. That being said, the percentages in Table 4 do provide us with a reasonable indication of how the three accounts perform. As is clear from the table, both the location-based and the parameter-based account clearly outperform the baseline. This is once again confirmed by the results of a t-test, presented in Table 5.

		t-value	p-value
location	parameters	-0.186	0.855
location	baseline	4.800	0.001
parameters	baseline	5.018	0.001

Table 5: Results of a two-tailed paired t-test for the two analyses vs. a frequency-based baseline

Both the location-based account and the parameter-based one differ significantly from our frequency-based baseline, thus suggesting that both provide plausible and credible accounts of the data set. Note, however, that in terms of the raw numbers listed in Table 4, the difference between the geographical and the grammatical analysis is no longer statistically significant. This strengthens our belief in the second part of our hypothesis, namely that there is a degree of complementarity in the type of data that is accounted for by both accounts. Exploring this intuition in more detail is a task we take up in the next section.

5 Visualizing the results

The kNN-experiments discussed in the previous section gave us a precise measure of how successful both analyses are in accounting for the variation data introduced in section 2, and how they fare with respect to a frequency-based baseline classification. On the other hand, these experiments did not provide any detailed or comparative information about where exactly the strengths and weaknesses of each account were located and to what extent they overlap or are complementary. In order to gain a clearer insight into this issue we decided to project the experimental results back onto a geographical map. Consider in this respect the maps in Figure 2.

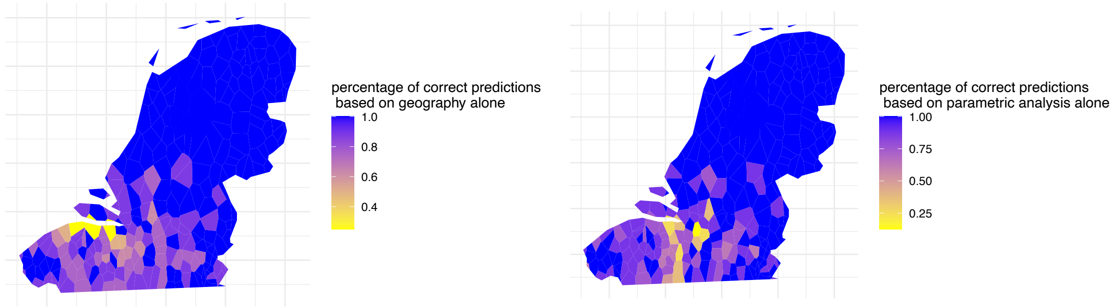

Figure 2: Percentage of correct predictions per dialect location

These maps represent our 267 dialect locations, color-coded according to the number of correct predictions each account made in this location.[9] In each location, we made eight predictions,[10] and the color indicates how many of those predictions were correct, ranging from 8 (solidly blue) to 1 (pure yellow). The map on the left-hand side visualizes the results of the location-only run of the experiment, while the map on the right-hand side shows the results of the parameters-only approach. As even a cursory inspection of these maps shows, the two accounts clearly differ in where they perform poorly (the yellow and orange areas). The location-based map on the left shows a clear bright spot in south of the province of Zeeland, specifically in the part of Zeeland that is contiguous with the Belgian province of West Flanders (an area known as Zeeuws Flanders). The country border here is indicative of a dialect split that is not reflected in the close geographical proximity with neigboring places in Belgium. In the parametric account this split is encoded in a difference in parameter setting: Zeeland has a negative setting for both the C- and the D-parameter, while West Flanders has a positive setting for both. The location-based account, however, has no access to this type of information, nor to the fact that there is a border separating neigboring places, and as a result it makes the wrong predictions.

The map on the right-hand side of Figure 2, though, has problems of its own. Note how there is brightly colored vertical line separating the province of East Flanders from the provinces of Antwerp and Flemish Brabant. In terms of the analysis, this represents the border between an area with a [$+AgrC$, $+D$, $+C$] parameter setting to one of type [$-AgrC$, $-D$, $+C$]. Given that the parameters in Van Craenenbroeck and van Koppen (2021) are categorical and binary, the transition between these two areas is expected to be sharp and abrupt. As is well-known from the traditional dialectological literature, however, the transition between the Flemish dialects in the west and the Brabant dialects in the east is a gradual one, with many border dialects showing characteristics from both dialect families. As is clear from the maps in Figure 2, a parametric account struggles with such a gradient transition, while a purely location-based one is much more successful, precisely because border dialects can take into account, i.e. be influenced by, properties of dialects from either side of the border.

Another way of visualizing these data is as in Figure 3. In these maps, we have split up the results per dialect phenomenon. This means that each location now represents a single prediction: does the phenomenon under consideration occur in this location or not? The colors now represent the predictions made by both accounts simulatenously: blue indicates that both accounts made the right prediction, while yellow means that both got it wrong. The orange color represents locations where the parameter-based account got it wrong, but the location-based one did not, and conversely, purple means that the parametric account made the right prediction, but the one based on geography did not. While it is hard to extract clear generalizations from such detailed and fine-grained maps, we do want to highlight two properties of these maps. One, in several maps we see the above-mentioned vertical border be-

[9] These maps are based on a kNN-experiment with $k = 3$, the first value where the location-only account did not differ significantly from the parameters-only account.

[10] Recall that we are not taking complementizer agreement and clitic doubling into consideration in the analysis, so as to not unfairly bias the results towards the parametric account. The absence of complementizer agreement also explains why the top of both maps is solidly blue: complementizer agreement is the only phenomenon from our data set that occurs in this area, and both accounts correctly predict none of the other phenomena occur there.

tween East Flanders on the one hand and Antwerp and Flemish Brabant on the other, and specifically in orange (meaning only the parametric account got it wrong) and in particular in the maps pertaining to the comparative marker (CMPR), the expletive (EXPL-T and ER-OBL), and the quirky V2-like imperatives (GO-GET). These are precisely the phenomena that were included in the C-parameter, but were not involved in setting that parameter, which might explain why the algorithm had problems correctly predicting them. The second observation we want to make concerns the relatively small role that the color yellow plays in these maps. With the exception of demonstrative doubling (THE-THAT) and perhaps quirky V2 (GO-GET), yellow seems to be a minority color. This suggests that the two accounts under consideration here indeed display a certain amount of complementarity, and that they are successful (and fail) in different contexts. Exploring these differences in more detail is something we will take up in future research.

Figure 3: Comparison of the predictions made by the two accounts per location

This concludes our comparison between the two accounts introduced in section 3. The one-sentence summary of our explorations is that both accounts represent credible and viable accounts of the variation data introduced in section 2, but that they seem to play complementary roles (and see Van Craenenbroeck et al. (2019) for comparable findings with respect to word order in verb clusters).

6 Summary & conclusions

This paper has been mostly methodological in nature. We have introduced a set of morphosyntactic variation data (in section 2) and two possible accounts of those data (in section 3), and we have then set out to compare those accounts (sections 4 and 5). Our primary goal, however, has not been to determine a winner between the two, but rather to introduce mechanisms and techniques that can be used to carry out such a comparison. Specifically, the k-nearest neighbor classification offers a useful and easily applicable means for carrying out such a comparison. We believe that formal-linguistic analyses can gain in strength and credibility by engaging in such comparisons and we look forward to continuing this line of research in the future.

References

Barbiers, Sjef, Johan van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet van der Ham. 2008. *Syntactische atlas van de Nederlandse dialecten. Deel II*. Amsterdam: Amsterdam University Press.

Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos, and Margreet van der Ham. 2005. *Syntactische atlas van de Nederlandse dialecten. Deel I*. Amsterdam: Amsterdam University Press.

Barbiers, Sjef, Marjo van Koppen, Hans Bennis, and Norbert Corver. 2016. Microcomparative MOrphosyntactic REsearch (MIMORE): Mapping partial grammars of Flemish, Brabantish and Dutch. *Lingua* 178:5--31.

Barbiers, Sjef, et al. 2006. *Dynamische syntactische atlas van de Nederlandse dialecten (dynasand)*. Meertens Institute. www.meertens.knaw.nl/sand/.

Belletti, Adriana. 2005. Extended doubling and the VP periphery. *Probus* 17:1--35.

Bennis, Hans. 1986. *Gaps and dummies*. Dordrecht: Foris Publications.

Brandner, Ellen. 2015. SynAlm: Tiefenbohrungen in einer Dialektlandschaft. In *Regionale Variation des Deutschen: Prokjekte und Perspektiven*, ed. Roland Kehrein, Alfred Lameli, and Stefan Rabamus, 289--322. Berlin: Mouton de Gruyter.

Buchelli, Claudia, and Elvira Glaser. 2002. The Syntactic Atlas of Swiss German Dialects: empirical and methodological problems. In *Syntactic microvariation*, ed. Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij, 41--74. Amsterdam: Meertens Institute.

Corver, Norbert, and Marjo van Koppen. 2018. Pronominalization and variation in Dutch demonstrative and possessive expressions. In *Atypical demonstratives*, ed. Marco Coniglio, Andrew Murphy, Eva Schlachter, and Tonjes Veenstra, 57--94. Berlin: Mouton de Gruyter. Ms. Utrecht University.

van Craenenbroeck, Jeroen. 2010. *The syntax of ellipsis. Evidence from Dutch dialects*. New York: OUP.

van Craenenbroeck, Jeroen. 2019. Expletives, locatives, and subject doubling. In *Linguistic variation: structure and interpretation*, ed. Ludovico Franco and Paolo Lorusso, 661--690. Berlin: Mouton de Gruyter.

van Craenenbroeck, Jeroen. 2020. Germanic specCP-expletives revisited: the view from Dutch microvariation. Ms. KU Leuven & Meertens Institute.

van Craenenbroeck, Jeroen, and Marjo van Koppen. 2002. Pronominal doubling and the structure of the left periphery in southern Dutch. In *Syntactic microvariation*, ed. Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij. http://www.meertens.knaw.nl/books/synmic/.

van Craenenbroeck, Jeroen, and Marjo van Koppen. 2008. Pronominal doubling in Dutch dialects: big DPs and coordinations. In *Microvariation in syntactic doubling.*, ed. Sjef Barbiers, Olaf Koeneman, Marika Lekakou, and Margreet van der Ham, volume 36 of *Syntax and Semantics*, 207--249. Bingley: Emerald.

van Craenenbroeck, Jeroen, and Marjo van Koppen. 2021. Microvariation and parameter hierarchies. Ms. KU Leuven & Meertens Institute & Utrecht University.

van Craenenbroeck, Jeroen, Marjo van Koppen, and Antal van den Bosch. 2019. A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters. *Language* 95:333--370.

Daelemans, W., and A. Van den Bosch. 2005. *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Fawcett, T. 2004. ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, Hewlett Packard Labs.

Fleischer, Jürg, Alexandra Lenz, and Helmut Weiß. 2015. Syntax hessischer Dialekte (SyHD). In *Regionale Variation des Deutschen: Prokjekte und Perspektiven*, ed. Roland Kehrein, Alfred Lameli, and Stefan Rabanus, 261--287. Berlin/Boston: Mouton de Gruyter.

Glaser, Elvira, and Gabriela Bart. 2015. Dialektsyntax des Schweizerdeutschen. In *Regionale Variation des Deutschen: Prokjekte und Perspektiven*, ed. Roland Kehrein, Alfred Lameli, and Stefan Rabanus, 79--105. Berlin: Mouton de Gruyter.

Greenacre, Michael. 2007. *Correspondence analysis in practice*. London & New York: Chapman & Hall, 2nd edition.

Grohmann, Kleanthes K. 2000. Prolific peripheries: a radical view from the left. Doctoral Dissertation, University of Maryland.

Haegeman, Liliane. 1986. *Er*-sentences in West-Flemish. Ms. Université de Genève.

Haegeman, Liliane. 1992. *Theory and description in generative syntax*. Cambridge: Cambridge University Press.

Haegeman, Liliane, and Anne Breitbarth. 2014. The distribution of preverbal *en* in (West) Flemish: syntactic and interpretive properties. *Lingua* 147:69--86.

Haegeman, Liliane, and Marjo van Koppen. 2012. Complementizer agreement and the relation between T and C. *Linguistic Inquiry* 43:441--454.

Kayne, Richard. 2005. Pronouns and their antecedents. In *Movement and silence*, 105--135. Oxford: Oxford University Press.

Klockmann, Heidi, Coppe van Urk, and Franca Wesseling. 2015. Agree is fallible, EPP is not: investigating EPP effects in Dutch. Handout of a talk at the Utrecht Syntax Interface Meetings.

van Koppen, Marjo. 2017. Complementizer agreement. In *The wiley-blackwell companion to syntax--second edition*, ed. Martin Everaert and Henk van Riemsdijk, 923--962. Wiley-Blackwell.

Laenzlinger, Christopher. 1998. *Comparative studies in word order variations: pronouns, adverbs and German clause structure*. Number 20 in Linguistics Today. Amsterdam: John Benjamins.

Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Lindstad, Arne Martinus, Anders Nøklestad, Janne Bondi Johannessen, and Øystein A. Vangsnes. 2009. The Nordic Dialect Database: Mapping microsyntactic variation in the Scandinavian languages. In *Nodalida 2009 conference proceedings*, ed. Kristiina Jokinen and Eckhard Bick, 283--286.

Nerbonne, John, and Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14:148--166.

Poletto, Cecilia. 2008. Doubling as a spare movement strategy. In *Microvariation in syntactic doubling*, ed. Sjef et al. Barbiers, volume 36 of *Syntax and Semantics*, 36--68. Bingley: Emerald.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Uriagereka, Juan. 1995. Aspects of the syntax of clitic placement in Western Romance. *Linguistic Inquiry* 26:79--124.

Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with s*. New York: Springer, fourth edition. URL https://www.stats.ox.ac.uk/pub/MASS4/, iSBN 0-387-95457-0.

de Vogelaer, Gunther. 2005. Subjectsmarkering in de Nederlandse en Friese dialecten. Doctoral Dissertation, Ghent University.

de Vos, Mark. 2006. Quirky verb-second in Afrikaans: complex predicates and head movement. In *Comparative studies in Germanic syntax: from Afrikaans to Zurich German*, ed. Jutta Hartmann and László Molnárfi, 89--114. Amsterdam: John Benjamins.

Zanuttini, Rafaella, Jim Wood, Jason Zentz, and Laurence Horn. 2018. The Yale Grammatical Diversity Project: Morphosyntactic variation in North American English. *Linguistics Vanguard* 4.

Zwart, Jan-Wouter. 1992. Dutch expletives and small clause predicate raising. In *Proceedings of North East Linguistic Society 22*, ed. K. Broderick, 477--491. Amherst, MA: GLSA.