

Parameters and Language Contact

Morphosyntactic Variation in Dutch Dialects

Jeroen van Craenenbroeck*

Meertens Institute/KU Leuven

`jeroen.van.craenenbroeck@meertens.knaw.nl`

ORCID: 0000-0003-4464-1689

Marjo van Koppen

Meertens Institute/Utrecht University

`marjo.van.koppen@meertens.knaw.nl`

ORCID: 0000-0002-9044-4312

October 14, 2022

Abstract

The central issue addressed in this paper is the formal linguistic notion of parameter as a predictor for the (non-)occurrence of multiple linguistic phenomena. We start from a parametric analysis of a microvariational data set and we introduce a way of gauging the success of that analysis. At the same time, we explore to what extent language contact—operationalized here as geographical proximity—can be seen as an explanatory factor that is complementary to the parametric analysis. Methodologically, the paper advocates the use of *k*-nearest neighbors classification as an interesting new technique in the linguist’s toolkit.

Keywords— microvariation, parameters, Dutch dialects, language contact

*We would like to thank the editors of this volume, Iván Ortega-Santos and Edita Gutiérrez Rodríguez, not only for their insightful comments and suggestions, but also for their patience and understanding. Many thanks also to the reviewers; their comments and questions have led to substantial improvements. The work reported on here is inspired by our earlier collaboration with Antal van den Bosch. The authors are listed in alphabetical order.

1 Introduction

The past twenty years have seen the birth of a 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 & Bart 2015, Buchelli & Glaser 2002), Alemannic (Brandner 2015), Hessian (Fleischer, Lenz & Weiß, 2015), and North American English (Zanuttini et al. 2018). The projects just mentioned differ from one another in e.g., the choice of linguistic phenomena that were focused on, the number of consultants and their sociolinguistic profile and the methodology used to elicit the data. There is one constant, however, that came out of 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 underlying assumption being that the syntactic properties of closely related dialect varieties are invariant (Brandner 2012, p. 118). 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 presents a serious challenge for generative approaches to language variation (see Brandner (2012) for general discussion). For instance, optionality and gradedness in the judgments abound in the data coming out of these dialect atlas projects, and it is not immediately clear how a parameter-based approach can handle such data: parameter settings are assumed to be categorical and binary, which leaves little to no room for gradability or optionality. This has prompted researchers to entertain other theoretical options such as the possibility of a single speaker having both settings of the same parameter, and thus effectively having multiple grammars (an idea going back to Kroch & Taylor (1997)). A second, related but distinct, issue concerns the question of whether a single parameter (setting) is responsible for the (non-)occurrence of multiple empirical phenomena. This idea has a venerable tradition going back at least to Rizzi's (1982, 1986) seminal work on the pro-drop parameter. Rizzi argued that a positive setting for this one parameter automatically leads to the occurrence within the language of a host of phenomena, including null referential and non-referential subjects, free subject-verb inversion, and a lack of *that*-trace effects. The strength of Rizzi's correlations was questioned early on (see e.g. Gilligan (1987) and Newmeyer (2005) for discussion), but the problems for this type of approach increase exponentially in microcomparative research: when the distribution of multiple empirical phenomena is examined not across a handful of languages but across literally hundreds of dialects of those languages, the chance of two phenomena having the exact same distribution effectively reduces to zero. That being said, the idea that one underlying parameter setting can have multiple surface effects remains prominent, even in the microcomparative approach. In an often-cited introduction to a volume on microcomparative syntax, Kayne (1996, p. xii) writes: "If it were possible to experiment on languages, a syntactician would construct an experiment of the following type: take a language, alter a single one of its observable syntactic properties, examine the result *to see what, if any, other property has changed as a consequence of the original manipulation*" (emphasis added). Once again, then, it is the correlation between empirical properties that is seen as a key characteristic of an underlying parameter setting.

The present paper takes this issue as its focal point: we start from a parametric analysis of a mi-

crovariational data set where the empirical correlations are not perfect and we introduce a way of gauging the success of that analysis. At the same time, we explore to what extent language contact—operationalized here as geographical proximity—provides an account for imperfect correlations and can thus be seen as an explanatory factor that is complementary to the parametric analysis. The paper is organized as follows: in section 2 we introduce the parametric analysis of Van Craenenbroeck & Van Koppen (2021) as well as the data it is based on. They propose three binary parameters to account for a data set containing 10 empirical phenomena in 260 dialects of Dutch, thus leading to eight different dialect types. In section 3 we test to what extent this analysis can correctly predict the empirical distribution of the 10 dialect phenomena by means of an experiment based on k -nearest neighbors classification (see also Van Craenenbroeck, Van Koppen & Van den Bosch 2019). We compare the results of this experiment to a classification based on geographical proximity, which we use as a proxy for language contact. The underlying idea is that increased language contact can lead to neighboring varieties sharing surface linguistic properties in spite of underlying grammatical differences.¹ Section 4 sums up and concludes.

2 The analysis of Van Craenenbroeck & Van Koppen (2021)

2.1 Introduction

This section introduces the main data and the analysis of Van Craenenbroeck & Van Koppen (2021) (henceforth VCVK). With respect to the analysis we focus on introducing those ingredients of the account that are relevant for the current discussion; for more details and further evidence in support of the analysis, we refer the reader to the original paper.

2.2 The data

The data that form the empirical basis of VCVK is a set of ten dialect phenomena found in the dialects of Dutch spoken in Belgium and the Netherlands. All of these data come from one of the above-mentioned dialect syntax projects, namely the Syntactic Atlas of Dutch Dialects (the SAND-project, Barbiers et al. (2005, 2008)). Data collection in the SAND-project proceeded in three rounds: a written questionnaire, a round of live oral interviews, and a round of telephone interviews. All data used in this paper come from the second and third round. Interviews were conducted in 267 dialect locations, with each interview involving at least two dialect speakers. Question types included translation questions, sentence completion tasks, and grammaticality judgments. In total, roughly 75 (morpho)syntactic variables were investigated, 10 of which were selected by VCVK for their parametric analysis. These ten phenomena are introduced and illustrated below. For a more detailed description of the design and methodology of the entire SAND-project, see Cornips & Jongenburger (2001).

The first phenomenon under investigation in VCVK is known as complementizer agreement (Van Koppen 2017), whereby a complementizer introducing a finite embedded clause can agree in person

¹In this paper we will not pursue the question of how to formally implement this type of variation: in terms of multiple grammars, variable grammars, etc. See Brandner (2012) for discussion and references.

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 & Van Koppen (2002); De Vogelaer (2005)). An example is given in (2).

- (2) Da gaan **ze** **zij** wel zegge.
 that go they_{clitic} they_{strong} PRT say
 ‘They will surely say that.’ Laarne, Barbiers et al. (2006)

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 in VCVK 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 & 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 in main clause-initial position. 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 in VCVK is the form of the comparative complementizer. Certain dialects (such as that of Oostkerke illustrated in (7)), use the coordinating disjunction *of* 'or' to introduce the standard of comparison, rather than the comparative marker *als* 'as'.

- (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 concerns the possibility of eliding the locative expletive pronoun *er* 'there' in embedded clauses and inverted main clauses (see Bennis (1986, p. 214), Zwart (1992), Klockmann, Van Urk, and Wesseling (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, p. 3)

Ninthly, some dialects allow for the combination of a definite determiner and a demonstrative in contexts of NP-ellipsis (Corver & Van Koppen 2018) as in (9), while others (like Standard Dutch) disallow 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 only 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, Barbiers et al. (2006)

The goal of this brief overview was not to examine these constructions in any detail—we refer to the literature mentioned above for in-depth descriptions—but rather to give the reader a general impression of the type of data analyzed by VCVK. The reason they select this specific set of phenomena is twofold: on the one hand, many of them have already been discussed and analyzed in the formal linguistic literature on Dutch dialects. This means that there are existing formal analyses of these phenomena, which in

turn makes it easier to identify any underlying parameters that might tie these phenomena together. The second reason for selecting this set of constructions concerns their geographical distribution. As shown in Figure 1, all ten of these phenomena are concentrated in the south west of the language area, which is another reason to think that there might be common grammatical principles underlying them.² At the same time, even a cursory glance at the maps in Figure 1 reveals that no two phenomena in this set have the exact same distribution. This makes VCVK's data and analysis highly relevant from the point of view of the current paper: they propose grammatical parameters whose (positive or negative) setting should be tracked by the combined occurrence or non-occurrence of multiple linguistic phenomena, but the expected correlations turn out not to be perfect.

After introducing their parametric account in the next subsection, we design an experiment in section 3 to investigate to what extent language contact—operationalized via geographic proximity—can account for some of those less than perfect correlations.

2.3 *The analysis*

VCVK present a parametric analysis of the data introduced in the previous section. Based in part on the geographical patterns shown in the maps in Figure 1³ and in part on the existing theoretical literature on these phenomena, they propose three grammatical parameters to account for the variation data. The first one sets complementizer agreement apart from all other phenomena, the second one combines clitic doubling with demonstrative doubling, and the third parameter unites the remaining seven constructions. We now proceed to discuss each parameter in turn.

The first one is what VCVK call the AgrC-parameter. It is formulated in (11).

- (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 & 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. In essence, then, what VCVK propose is that dialects that have a positive setting for this parameter feature an AgrC-projection in their left periphery as proposed by Shlonsky (1994). More broadly, the type of variation displayed by these dialects fits nicely

²The 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. The parenthesized numbers following the abbreviations refer to the number of dialects where the phenomenon in question is attested.

³More specifically, VCVK perform a Correspondence Analysis (Greenacre 2007) on the raw data underlying the maps in Figure 1 to determine which phenomena should be grouped together. As this quantitative part of their analysis is orthogonal to the concerns of the present paper, we leave it undiscussed in the remainder of this paper and refer the reader to the original paper for details.



Figure 1: Geographical distribution of the ten dialect phenomena (Barbiers *et al.*, 2006)

into Roberts & Holmberg's (2010, p. 49) discussion of null arguments. Roberts & Holmberg propose that the cross-linguistic distribution of null arguments can be captured by positing phi-probes on (sets of) specific syntactic heads. An indication that it makes sense to analyze complementizer agreement along the same lines is the fact that this phenomenon sometimes licenses null subject arguments as well:

- (12) Et geberde doest fot giest.
 it happened when.2SG away went
 'It happened when you went away.'

Lies, Barbiers *et al.* (2006)

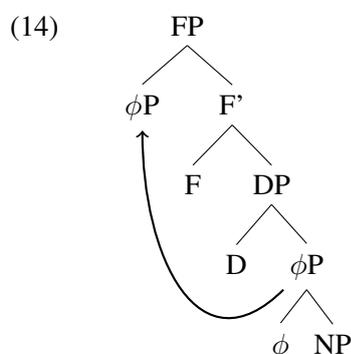
One thing to note about VCVK's AgrC-parameter in the context of this paper is that it sets one phenomenon, complementizer agreement, apart from all the others. In other words, VCVK hypothesize that the distribution of complementizer agreement is unconnected and hence orthogonal to that of the remaining nine phenomena introduced in section 2.2. The second parameter, however, does link up two

constructions. It is formulated in (13).

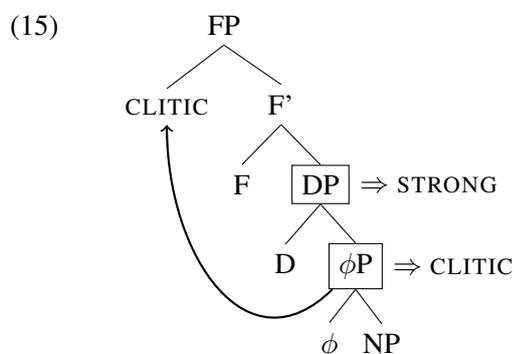
(13) **D-parameter:**

DP {does/does not} have a split left periphery.

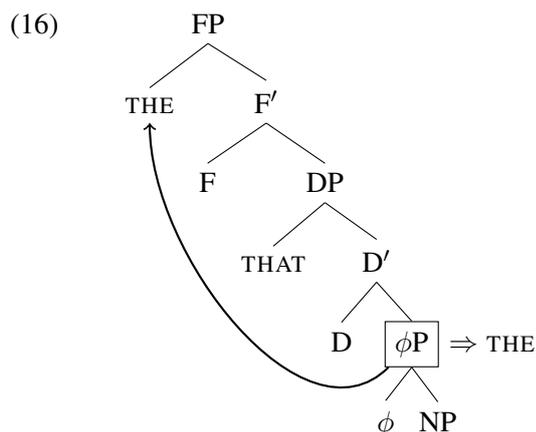
The idea here is that certain dialects have an extended left periphery in their nominal domain, while others do not. This additional structural space can serve as a landing site for DP-internal movement operations, thus making it possible for certain constructions to arise that would be disallowed in 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 & Van Koppen (2008), they argue that the clitic is the result of DP-internal movement of ϕ P into the extended left periphery of the DP, as illustrated in (14).



In this structure, ϕ P moves into the left periphery of the DP. Note that it cannot move to specDP, as that would constitute an anti-locality violation (Abels (2003)). In dialects with a positive setting for the D-parameter in (13), however, there is additional structural space in the left periphery (indicated as FP here) that can serve as a landing site for ϕ P-movement. Given that this movement is a prerequisite for clitic doubling—see the spell-out of the structure in (15) as sketched in (15)—this means that only dialects with a positive setting for the D-parameter can feature clitic doubling.



A similar line of reasoning applies to determiner-demonstrative doubling: Barbiers et al. (2016) argue on independent grounds that what looks like a determiner is actually the spell-out of a ϕ P that has raised into the extended left periphery of the DP, as shown in (16).



Once again, there is a structural condition that has to be met in order for this construction to be able to arise: the left periphery of the DP has to be ‘rich’ or elaborate enough to host the ϕ P that will be spelled out as *de*, and this richness is only found in dialects that have a positive setting for the D-parameter.

Summing up, VCVK assume these two phenomena to be regulated by a single parameter, the one in (13). The null hypothesis, then, is that the geographical distribution of these two phenomena should 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 the comparison between the relevant maps in Figure 1 shows, however, this null hypothesis is not met. As pointed out above, this discrepancy is precisely the topic that this paper focuses on and that will be the subject of the experiment in the next section, but for VCVK, it raises the practical question of how to determine the setting of the D-parameter in a specific dialect: 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 propose, 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.

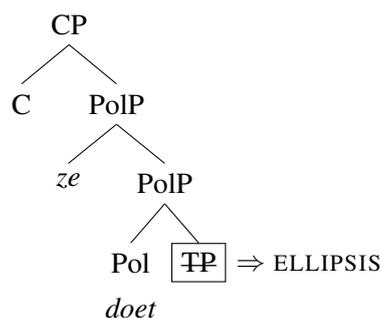
The third and final parameter that VCVK propose bundles together the remaining seven empirical phenomena listed in subsection 2.2. It mirrors the D-parameter, but applies at the clausal level:

(17) **C-parameter**

CP {does/does not} have a split left periphery.

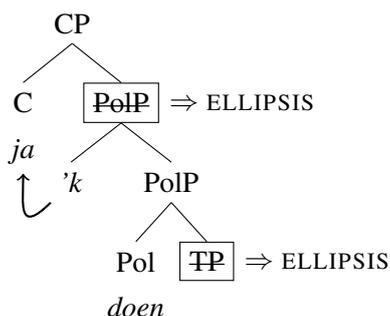
The reasoning here parallels the one outlined above with respect to the D-parameter: dialects with a positive setting for the C-parameter provide more structural space in their clausal left periphery, and so constructions that specifically target or make use of this extra structural space are sensitive to the setting of this parameter. With respect to short *do*-replies, Van Craenenbroeck (2010) argues that the verb *doen* ‘to do’ spells out a high left-peripheral polarity phrase (PolP), licensing ellipsis of TP:

(18)



In other words, the occurrence of short *do*-replies is dependent on the presence of a high left-peripheral PolP, and this projection is only present in dialects that have a positive setting for the parameter in (17). Moreover, two other phenomena are also directly dependent on this same projection according to Van Craenenbroeck (2010): on the one hand, the negative clitic *en* is a spell-out of Pol⁰—see also Haegeman & Breitbarth (2014) for discussion—and on the other, the occurrence of clitics on ‘yes’ and ‘no’ underlyingly derives from short *do*-replies and features an additional ellipsis process on top of the structure in (18):

(19)



With respect to the use of *t* ‘it’ as an expletive and the obligatory versus optional occurrence of the locative expletive in inversion and embedded clauses, VCVK follow Van Craenenbroeck (2020), who argues both these phenomena to be a reflex of a split versus an unsplit CP-domain. He proposes that the *t*-expletive is not a pronoun, but rather a main clause complementizer spelling out the Force-head. Given that the verb also needs a C-head to move to—all the dialects under discussion are Verb Second—this configuration is only found in dialects that feature (at least) two separate heads in their left periphery (one to host the fronted verb and one that is spelled out as the main clause complementizer), i.e. in dialects that have a positive setting for the C-parameter in (17). The optional deletion of the locative expletive on the other hand, Van Craenenbroeck (2020) assumes to be the reflex of another locative expression moving into the canonical subject position, following ideas put forward by Klockmann, Van Urk & Wesseling (2015). Once again, this is an option that 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, the only two phenomena in the original set of ten for which there was no pre-existing formal-theoretical analysis for VCVK to fall back on. With respect to the former, they suggest that *come* and *go* have been grammaticalized and spell out a functional head higher than the one hosting the finite (imperative) verb. In so doing, they follow McCloskey’s (1997, p. 214) analysis of Ulster English

examples like the one in (20). McCloskey argues that the element *gon* is an imperative marker which spells out a head in the CP-domain. Given that the imperative verb in the dialect Dutch examples also occupies a left-peripheral head, the existence of the *come/go get*-construction shows that there need to be at least two separate C-heads. This means that this construction only occurs in dialects that have a positive setting for the C-parameter in (17).

(20) Gon make us (you) a cup of tea.

Ulster English

With regard to the use of *of* ‘or’ as a standard marker, VCVK take it to be indicative of the degree of syncretism in the complementizer inventory of a dialect: when dialects use *of* ‘or’ as a standard marker, they have a different complementizer for comparative and conditional clauses, whereas dialects that do not use *of* ‘or’ typically have the same complementizer in these two clause types, namely (a variant of) *als* ‘as’. VCVK take this (lack of) morphological richness to be indicative of the underlying syntactic richness in the left periphery: different complementizers in comparative and conditional clauses means separate left peripheral projections for these two clause types, whereas one and the same complementizer implies a single, unsplit projection for the two types (see Bobaljik & Thráinsson (1998) for a comparable line of reasoning with respect to the TP-domain). More generally, this means that dialects that use *of* ‘or’ as a comparative complementizer have a positive setting for the C-parameter in (17) according to VCVK.

Summing up, 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 (17). 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 for VCVK of how to determine the setting of this parameter in a specific dialect. They 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, and/or clitics on ‘yes’ and ‘no’.

This concludes our overview of VCVK’s parametric account of the morphosyntactic variation introduced in section 2.2. It reduces said variation to the interplay between three parameters: the AgrC-parameter in (11), the D-parameter in (13), and the C-parameter in (17). The first one specifically targets complementizer agreement, the second one bundles together clitic doubling and determiner-demonstrative doubling, and the third one covers the remaining seven phenomena. Together, these three binary parameters yield eight different dialect types or groups. The size and geographical distribution of these groups is represented in Figure 2. Membership of a group is determined by the presence of complementizer agreement (for the AgrC-parameter), the presence of clitic doubling (for the D-parameter), and the presence of at least one of the three polarity-related constructions (for the C-parameter).

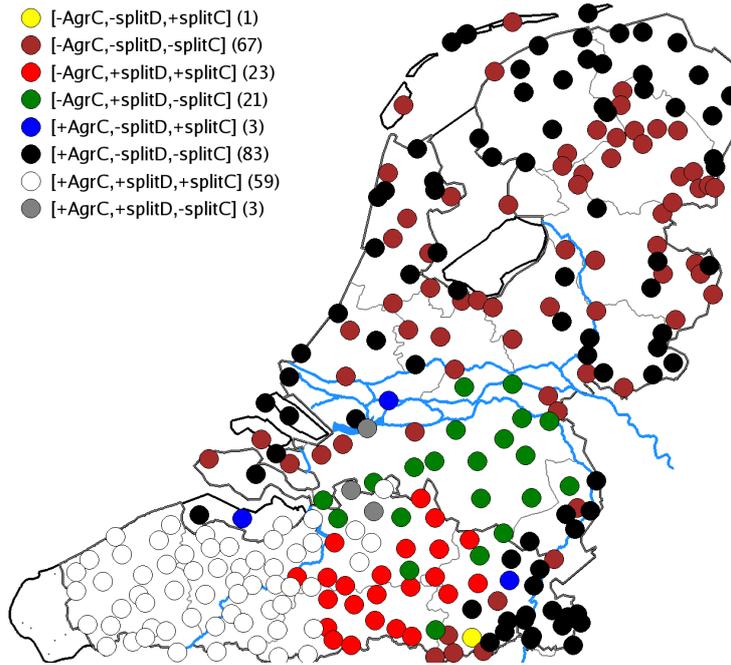


Figure 2: Eight dialect types as predicted by VCVK’s parametric account

3 The experiment: the role of geographical proximity

3.1 Introduction

In this section we take the parametric account outlined in the previous section as the starting point for an experiment based on k -nearest neighbors classification in order to gauge the success of VCVK’s analysis, specifically when it comes to the predicted correlations between the various phenomena and the degree to which language contact—operationalized through geographical proximity—can serve as an additional or complementary explanatory factor. The section is organized as follows. In subsection 3.2 we introduce the method and lay out the experiment, and in subsection 3.3 we present, discuss, and visualize the results.

3.2 k -nearest neighbors classification

The k -nearest neighbors (henceforth k NN) classifier is a form of supervised machine learning (Daelemans & Van den Bosch 2005).⁴ The model takes as its input a set of labeled training data, and then uses this information to make predictions about (the classification of) unseen data points. Let us use a simple example to explain the basic workings of k NN. Consider the plot in Figure 3.

Suppose we use the colored points (the red triangles and blue circles) as our training data. More specifically, we would feed into the model a set of eight data points consisting of a set of coordinates and a classification (red triangle or blue circle). Then, we can present the model with a new data point,

⁴All k NN-calculations reported on in this paper were carried using the Tilburg Memory-Based Learner (TiMBL), see Daelemans et al. (2018) for detailed documentation.

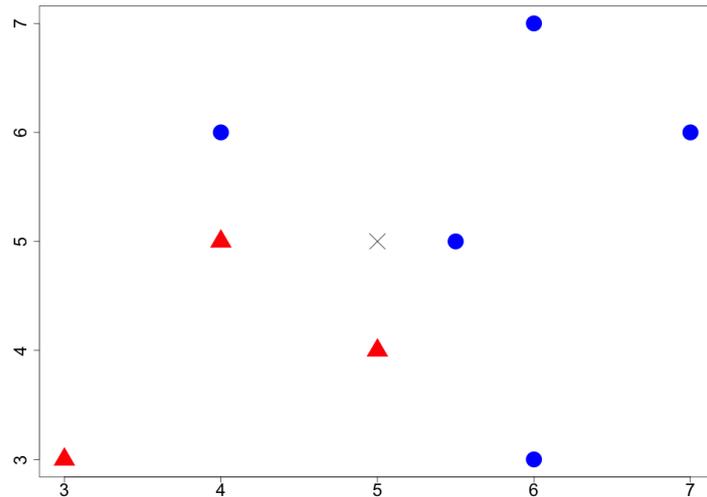


Figure 3: Basic illustration of k NN

represented by the black cross in Figure 3, and ask what its classification should be: blue circle or red triangle? In order to answer this question, the k NN-algorithm looks at the nearest neighbors of the new data point and assumes the new point will have the same classification as (the majority of) those neighbors. The hyperparameter k indicates the number of neighbors taken into consideration. For example, if $k = 1$ then the classifier looks only at the single closest neighbor, which in this case would be the blue circle at coordinates (5.5,5). Accordingly, the new data point would be classified as a blue circle. If $k = 3$, however, the set of closest neighbors would consist of two red triangles and one blue circle, and the new point would be classified as a red triangle.⁵

Turning from blue circles and red triangles to linguistic data, it is fairly straightforward to see how we can apply the mechanism just sketched to the geographical distribution of the dialect phenomena under discussion in this paper. Consider in Figure 4 the distribution of determiner-demonstrative doubling. On this map, black circles represent places where determiner-demonstrative doubling is attested, and transparent circles are dialects where the phenomenon is absent. When we feed this information into a k NN-model, we can then use it to make predictions about the presence or absence of determiner-demonstrative doubling in other dialect locations. However, given that we do not have any information about the occurrence of determiner-demonstrative doubling beyond the 260 dialects in VCVK’s data set and so would not be able to test the correctness of the k NN-predictions, we adopt a specific implementation of the k NN-classifier in this paper called ‘leave-one-out’. In this approach, each of the 260 dialects in turn acts as the new, unseen data point, while the remaining 259 dialects serve as the known, memorized training data set. For each of the ten phenomena, the k NN-classifier will then make 260 individual predictions, which we can classify as true or false, and which we can then apply some basic statistics on (see below).

⁵In the case of a tie—which can arise when k is even—various implementations are possible. For example, the algorithm can randomly choose a classification, or it can choose the one that is more frequent in the entire data set, etc.



Figure 4: The distribution of determiner-demonstrative doubling across Dutch dialects

What is interesting about k NN-classification in the context of the present paper is the fact that it can use different types of distance measures in order to determine the nearest neighbors. So far, we have only looked at one such measure, namely Euclidean distance. For one half of our research question, this makes perfect sense: if we want to know whether the distribution of, say, clitic doubling is influenced by language contact, we can determine to what extent the presence or absence of clitic doubling in a specific dialect location can be predicted based on the distribution of this phenomenon in neighboring places, the idea being that speakers of dialects that are in close geographical proximity will have more contact—and hence more contact-induced language variation—than speakers that are geographically distant. At the same time, however, we also want to evaluate VCVK’s parametric account of the variation, and in their account, proximity is measured not in terms of geography, but in terms of parameter settings. Consider the hypothetical data set in (21).

(21)	AgrC	splitD	splitC	clitic doubling
dialect A	yes	yes	no	yes
dialect B	yes	yes	no	yes
dialect C	yes	no	no	no
dialect D	no	yes	no	yes
dialect E	no	no	yes	yes

Once again, we are interested in predicting whether or not a dialect features a certain construction—i.e. the values in the final column in the table in (21)—and once again, we want to do so by looking at whether or not that construction occurs in dialects that are similar to the dialect under consideration.

The only difference is that in this case similarity is not determined by geographical distance, but by parameter settings. In the table in (21), dialects A and B are maximally close to one another, in that they have the exact same parameter settings. Dialects C and D are slightly less close—they both differ in the value of one parameter—but both of them are closer to A and B than E is, which has a different value for all three parameters. In this way, we can also use k NN-based classification to determine how successful VCVK’s parametric analysis is in describing the data set introduced in section 2.2. One important difference, though, compared to k NN based on Euclidean distance concerns the interpretation of k . When it is set to 1, we are looking for the unique *parameter setting* that is maximally close to that of the dialect under consideration, not the unique *dialect*. In practice, many different dialects can have the same parameter setting (see the dialect groups in Figure 2). This means that all those dialects will be taken into consideration as nearest neighbors and the prediction will be based on whether the construction occurs in the majority of those dialects. Put differently, in this version of the algorithm, the value of k refers to k -nearest *distances* rather than k -nearest examples (Daelemans et al. 2018, p. 25).⁶

With this much as background, we can now proceed to describe the experiment we carried out in more detail. As mentioned above, we opted for k NN-classification of the type ‘leave one out’, whereby each of the 260 dialects in turn acts as the unknown, to-be-predicted value, while the remaining 259 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 2.2, 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. We initially set the value of k to 1, but will also report the results of experiments with higher values for k in the next subsection. The outcome of (one run of) one experiment is a set of 260 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. 260, but this is too crude a measure for cases where a phenomenon is rare. For instance, the use of *of* ‘or’ as a standard marker only occurs in 33 of the 260 dialects. A model that always predicts this phenomenon to be absent would still get 227 of the 260 cases correct, an accuracy of 87%, in spite of the fact that none of the positive cases were correctly identified. In order to counter this effect, we compute 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. This is the measure we will use to gauge the success of the various experimental runs in predicting the variation data.

⁶This also increases the chances of a tie (cf. fn5). TiMBL deals with ties as follows: “First, the value of the k parameter is incremented by 1, and the additional nearest neighbors at this new k th distance are added to the current nearest neighbor set (k is subsequently reset to its user-specified value). If the tie in the class distribution persists, then the class label is selected with the highest overall occurrence in the training set. If that is also equal, then the first class is taken that was encountered when reading the training instance file.” (Daelemans et al. 2018, p. 32)

3.3 Results of the experiment

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.⁷

	location	parameters	location+parameters
CA	0.733	1.000	1.000
CD	0.927	1.000	1.000
SDR	0.891	0.938	0.905
CYN	0.927	0.901	0.935
NEG	0.902	0.957	0.936
COMPR	0.846	0.943	0.848
EXPL-T	0.920	0.907	0.924
GO-GET	0.788	0.919	0.779
THE-THAT	0.824	0.868	0.827
ER-OBL	0.836	0.959	0.853
AVERAGE	0.867	0.924	0.876
SD	0.050	0.031	0.058

Table 1: AUC-values of the k NN-experiment

When discussing the results, we set aside those pertaining to complementizer agreement (CA) and clitic doubling (CD), i.e. the greyed-out rows in Table 1. Recall from subsection 2.3 that VCVK use 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 parameter value.⁸ Overall, it seems clear that a classification based on the grammatical parameters of Van Craenenbroeck & Van Koppen (2021) fares better than a geography-based one: in six out of

⁷Recall 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.

⁸Note 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.

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 (COMPR) or the obligatory nature of the *there*-expletive (ER-OBL).⁹ This general intuition is confirmed by the results of a two-tailed paired *t*-test, a summary of which is given in Table 2.

		<i>t</i> -value	<i>p</i> -value
location	parameters	-2.734	0.018
parameters	location+parameters	2.072	0.063
location	location+parameters	-0.337	0.741

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 VCVK is significantly different ($p = 0.018$) from one that only uses geographical information (longitude and latitude), and that neither account differs significantly from one that uses a combination of both types of information ($p > 0.05$ in both cases). On the whole, this bodes well for the parametric analysis of VCVK: it seems to be able to capture a substantial portion of the attested variation. At the same time, we should not be so quick to dismiss the role played by language contact, i.e. by geographical proximity as encoded in the geography-based run of the experiment. Recall that we started out by setting the value of the hyperparameter k to 1. For the parameter-based account, this is a reasonable setting, as it means that the value for a particular dialect is determined based on those dialects that have the exact same parameter setting. Groups defined by specific parameter settings are thus considered to be natural classes and it is those classes that the k NN-algorithm is taking into account. Increasing the value of k is expected to decrease the accuracy of the parameters-only account, as it would involve looking at dialects with increasingly different parameter settings to predict the linguistic properties of the dialect under consideration.

The opposite prediction holds, however, for the geography-based account: looking at only one neighboring dialect (the closest one) seems to be too crude a measure for determining the linguistic profile of a dialect, and increasing the number of neighbors taken into consideration should increase the accuracy of the approach. This is indeed what we see in Table 3, where we report on experimental runs with k ranging from 1 to 10.

As expected, the parameters-only approach quickly deteriorates as the value of k goes up. It bottoms out at chance behavior for values of $k = 4$ and up, given that at this point it takes all dialects into consideration when making a prediction and always chooses the value that is most frequent in the entire data set. The location-based approach on the other hand, does benefit from raising the value of k : taking more neighboring dialects into account leads to better predictions, with peak performance at $k = 8$.¹⁰

⁹Recall from subsection 2.3 that the C-parameter is set based on the values of the three polarity-related phenomena only.

¹⁰The fact that the average AUC-value for even values of k is identical to that of $k + 1$ is due to the way ties are resolved, as described in fn6: if there is a tie for $k = n$ (which can only arise when n is even) the algorithm

k	average AUC location	average AUC parameters	t -value	p -value
1	0.867	0.924	-2.734	0.018
2	0.880	0.708	2.097	0.071
3	0.880	0.507	19.698	0.000
4	0.891	0.500	21.800	0.000
5	0.891	0.500	21.800	0.000
6	0.896	0.500	25.479	0.000
7	0.896	0.500	25.479	0.000
8	0.902	0.500	37.382	0.000
9	0.902	0.500	37.382	0.000
10	0.899	0.500	38.932	0.000

Table 3: Results of experimental runs with different values for k

In short, both the parametric account of VCVK and a contact-based approach seem to have their merits in describing the variation attested in this data set. This might lead one to hypothesize that both accounts meet a certain basic threshold in terms of their ability to model the data—i.e. that both provide a reasonable account for the variation data introduced in section 2.2—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 other words, maybe the areas where the parametric analysis makes the wrong predictions—in particular because phenomena that are assumed to be dependent on the same parameter do not pattern alike, see the discussion in section 2—are the ones where the contact-based approach fares better and vice versa. In order to test this hypothesis, we can project the experimental results back onto a geographical map. Consider in this respect the maps in Figure 5.

These maps represent the 260 dialect locations color-coded according to the number of correct predictions each account made in the relevant location. These maps are based on a k NN-experiment with $k = 8$ for the location-based approach and $k = 1$ for the parameter-based one, i.e. in both cases we have selected the value for k that yields the best predictions. In each location, we made eight predictions,¹¹ and the color indicates how many of those predictions were correct, ranging from 7 or 8 (white) over 5 or 6 (yellow) and 3 or 4 (red) to 1 or 2 (dark red). 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 is clear from visually inspecting the maps, the two accounts differ

increases the value of k by 1, and if there is no tie, that means that at least $\frac{n}{2} + 1$ dialects have the same value, which will yield the same winner in case $k = n + 1$ (because $\frac{n}{2} + 1 > \frac{n+1}{2}$).

¹¹Recall 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 white: 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.

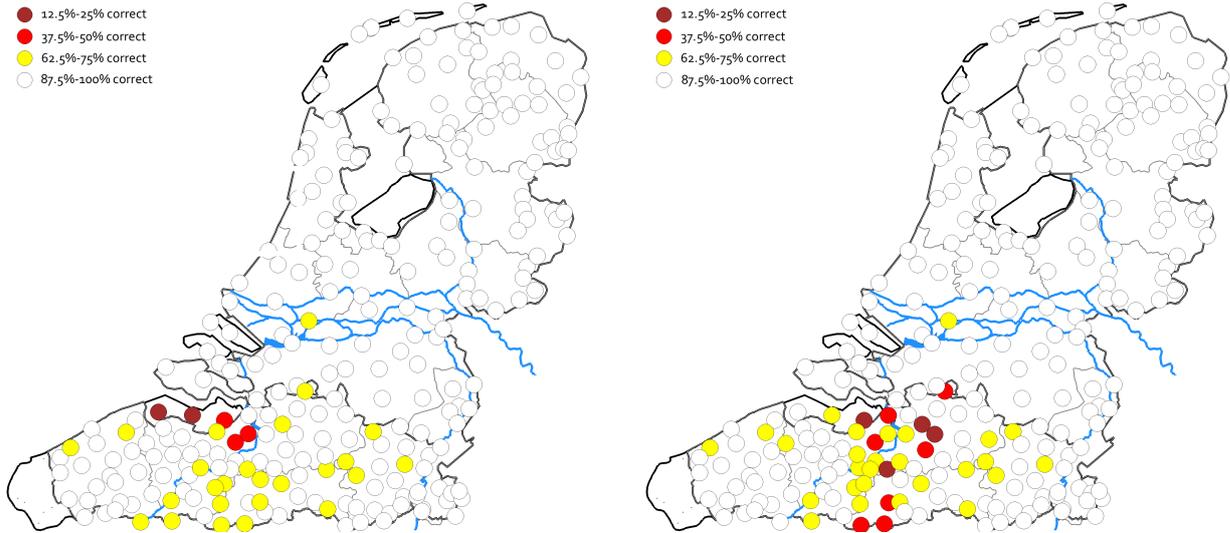


Figure 5: Percentage of correct predictions per dialect location for the location-only account (left) and the parameters-only account (right)

in where they perform poorly (the red and dark red areas). The location-based map on the left shows a clear red zone in the 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 neighboring places in Belgium. In other words, this is a part of the language area where geographical proximity is arguably not a great indicator of language contact, given the presence of a country border. This is also precisely the type of context where the parametric account fares better. It encodes the split between (the provinces of) Belgium and (those of) the Netherlands 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 map on the right-hand side of Figure 5 shows weak predictions in an entirely different area, namely the one separating the province of East Flanders from the provinces of Antwerp and Flemish Brabant. In terms of VCVK's analysis, this represents the border between an area with a $[+AgrC, +splitD, +splitC]$ parameter setting to one of type $[-AgrC, -splitD, +splitC]$. Given that the parameters in VCVK's account are categorical and binary, the transition between these two areas is expected to be sharp and abrupt. As is well-known from the dialectological literature, however, the border between the Flemish dialects in the south-west of the Belgian dialect area and the Brabant dialects in the middle of that area is a gradual one, with many border dialects showing characteristics from both dialect families (Taeldeman 2000). As is clear from the maps in Figure 5, this is an area where language contact provides a better account of the attested variation than a strict parametric analysis.

4 Summary & conclusions

The central issue addressed in this paper is the notion of a parameter as a predictor for the occurrence or non-occurrence of multiple linguistic phenomena, especially in an era where the amount of microvariation data has increased considerably and perfect correlations are rarely attested, if at all. We have introduced a set of morphosyntactic variation data and a parameter-based analysis of those data, and we have presented a way to not only gauge the success of such an analysis in accounting for the variation, but also to explore to what extent language contact can serve as an additional or alternative explanatory factor. The idea is that contact-induced language variation can soften or blur the hard grammar transitions expected by a strict parametric account.

Methodologically, we have advocated for the use of k -nearest neighbors classification as an interesting and novel method for evaluating and comparing analyses. We believe that formal-linguistic analyses can gain in strength and credibility by being subjected to these types of evaluations, and we look forward to continuing this line of research in the future.

References

- Abels, K. (2003), Successive cyclicity, anti-locality, and adposition stranding, PhD thesis, University of Connecticut at Storrs.
- Barbiers, S., Auwera, J. v. d., Bennis, H., Boef, E., Vogelaer, G. D. & Ham, M. v. d. (2008), *Syntactische atlas van de Nederlandse dialecten. Deel II*, Amsterdam University Press, Amsterdam.
- Barbiers, S., Bennis, H., Vogelaer, G. D., Devos, M. & Ham, M. v. d. (2005), *Syntactische atlas van de Nederlandse dialecten. Deel I*, Amsterdam University Press, Amsterdam.
- Barbiers, S., van Koppen, M., Bennis, H. & Corver, N. (2016), ‘Microcomparative MORphosyntactic REsearch (MIMORE): Mapping partial grammars of Flemish, Brabantish and Dutch’, *Lingua* **178**, 5–31.
- Barbiers, S. et al. (2006), *Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND)*, Meertens Institute. www.meertens.knaw.nl/sand/.
- Belletti, A. (2005), ‘Extended doubling and the VP periphery’, *Probus* **17**(1), 1–35.
- Bennis, H. (1986), *Gaps and dummies*, Foris Publications, Dordrecht.
- Bobaljik, J. & Thráinsson, H. (1998), ‘Two heads aren’t always better than one’, *Syntax* **1**, 37–98.
- Brandner, E. (2012), ‘Syntactic microvariation’, *Language and Linguistics Compass* **6**(2), 113–130.
- Brandner, E. (2015), SynAlm: Tiefenbohrungen in einer Dialektlandschaft, in R. Kehrein, A. Lameli & S. Rabamus, eds, ‘Regionale Variation des Deutschen: Prokjekte und Perspektiven’, Mouton de Gruyter, Berlin, pp. 289–322.
- Buchelli, C. & Glaser, E. (2002), The Syntactic Atlas of Swiss German Dialects: empirical and methodological problems, in S. Barbiers, L. Cornips & S. van der Kleij, eds, ‘Syntactic microvariation’, Meertens Institute, Amsterdam, pp. 41–74.
- Cornips, L. & Jongenburger, W. (2001), ‘Elicitation techniques in a Dutch syntactic dialect atlas project’, *Linguistics in the Netherlands* **18**, 53–63.
- Corver, N. & van Koppen, M. (2018), Pronominalization and variation in Dutch demonstrative and possessive expressions, in M. Coniglio, A. Murphy, E. Schlachter & T. Veenstra, eds, ‘Atypical demonstratives’, Mouton de Gruyter, Berlin, pp. 57–94.
- Daelemans, W. & Van den Bosch, A. (2005), *Memory-based language processing*, Cambridge University Press, Cambridge, UK.
- Daelemans, W., Zavrel, J., van der Sloot, K. & van den Bosch, A. (2018), *TiMBL: Tilburg Memory-Based Learner*, Tilburg University & University of Antwerp, Tilburg/Antwerp.

- de Vogelaer, G. (2005), *Subjectsmarkering in de Nederlandse en Friese dialecten*, PhD thesis, Ghent University.
- de Vos, M. (2006), Quirky verb-second in Afrikaans: complex predicates and head movement, in J. Hartmann & L. Molnárfi, eds, 'Comparative Studies in Germanic Syntax: from Afrikaans to Zurich German', John Benjamins, Amsterdam, pp. 89–114.
- Fawcett, T. (2004), ROC graphs: Notes and practical considerations for researchers, Technical Report HPL-2003-4, Hewlett Packard Labs.
- Fleischer, J., Lenz, A. & Weiß, H. (2015), Syntax hessischer Dialekte (SyHD), in R. Kehrein, A. Lameli & S. Rabanus, eds, 'Regionale Variation des Deutschen: Prokjekte und Perspektiven', Mouton de Gruyter, Berlin/Boston, pp. 261–287.
- Gilligan, G. (1987), *A cross-linguistic approach to the pro-drop parameter*, PhD thesis, USC.
- Glaser, E. & Bart, G. (2015), Dialektsyntax des Schweizerdeutschen, in R. Kehrein, A. Lameli & S. Rabanus, eds, 'Regionale Variation des Deutschen: Prokjekte und Perspektiven', Mouton de Gruyter, Berlin, pp. 79–105.
- Greenacre, M. (2007), *Correspondence analysis in practice*, 2nd edn, Chapman & Hall, London & New York.
- Grohmann, K. K. (2000), *Prolific peripheries: a radical view from the left*, PhD thesis, University of Maryland.
- Haegeman, L. (1986), *Er-sentences in West-Flemish*. Ms. Université de Genève.
- Haegeman, L. (1992), *Theory and Description in Generative Syntax*, Cambridge University Press, Cambridge.
- Haegeman, L. & Breitbarth, A. (2014), 'The distribution of preverbal *en* in (West) Flemish: syntactic and interpretive properties', *Lingua* **147**, 69–86.
- Haegeman, L. & van Koppen, M. (2012), 'Complementizer agreement and the relation between T and C', *Linguistic Inquiry* **43**(3), 441–454.
- Kayne, R. (1996), Microparametric syntax: some introductory remarks, in J. Black & V. Motapanyane, eds, 'Microparametric syntax and dialect variation', John Benjamins, Amsterdam, pp. ix–xviii.
- Kayne, R. (2005), Pronouns and their antecedents, in 'Movement and Silence', Oxford University Press, Oxford, pp. 105–135.
- Klockmann, H., van Urk, C. & Wesseling, F. (2015), Agree is fallible, EPP is not: investigating EPP effects in Dutch. Handout of a talk at the Utrecht Syntax Interface Meetings.

- Kroch, A. & Taylor, A. (1997), Verb movement in Old and Middle English: dialect variation and language contact, in A. van Kemenade & N. Vincent, eds, 'Parameters of morphosyntactic change', Cambridge University Press, Cambridge, pp. 298–325.
- Laenzlinger, C. (1998), *Comparative studies in word order variations: pronouns, adverbs and German clause structure*, number 20 in 'Linguistics Today', Amsterdam: John Benjamins.
- Lindstad, A. M., Nøklestad, A., Johannessen, J. B. & Vangsnes, Ø. A. (2009), The Nordic Dialect Database: Mapping microsyntactic variation in the Scandinavian languages, in K. Jokinen & E. Bick, eds, 'NODALIDA 2009 Conference Proceedings', pp. 283–286.
- McCloskey, J. (1997), Subjecthood and subject positions, in L. Haegeman, ed., 'Elements of grammar', Kluwer Academic Publishers, Dordrecht, pp. 197–235.
- Newmeyer, F. J. (2005), Against a parameter-setting approach to language variation, in P. Pica, J. Rooryck & J. van Craenenbroeck, eds, 'Linguistic Variation Yearbook', Benjamins, Amsterdam, pp. 181–234.
- Poletto, C. (2008), Doubling as a spare movement strategy, in S. e. a. Barbiers, ed., 'Microvariation in syntactic doubling', Vol. 36 of *Syntax and Semantics*, Bingley: Emerald, pp. 36–68.
- Rizzi, L. (1982), *Issues in Italian syntax*, Foris Publications, Dordrecht, Holland.
- Rizzi, L. (1986), 'Null objects in Italian and the theory of *pro*', *Linguistic Inquiry* 17(3), 501–557.
- Roberts, I. & Holmberg, A. (2010), Introduction: parameters in minimalist theory, in T. Biberauer, A. Holmberg, I. Roberts & M. Sheehan, eds, 'Parametric variation. Null subjects in minimalist theory', Cambridge University Press, Cambridge, pp. 1–57.
- Shlonsky, U. (1994), 'Agreement in Comp', *The Linguistic Review* 11(3-4), 351–376.
- Taeldeman, J. (2000), De regenboog van de Vlaamse dialecten, in 'Het taallandschap in Vlaanderen', Vakgroep Nederlandse taalkunde Gent, Gent, pp. 1–15.
- Uriagereka, J. (1995), 'Aspects of the syntax of clitic placement in Western Romance', *Linguistic Inquiry* 26(1), 79–124.
- van Craenenbroeck, J. (2010), *The syntax of ellipsis. Evidence from Dutch dialects*, OUP, New York.
- van Craenenbroeck, J. (2019), Expletives, locatives, and subject doubling, in L. Franco & P. Lorusso, eds, 'Linguistic variation: structure and interpretation', Mouton de Gruyter, Berlin, pp. 661–690.
- van Craenenbroeck, J. (2020), Germanic specCP-expletives revisited: the view from Dutch microvariation. Ms. KU Leuven & Meertens Institute.
- van Craenenbroeck, J. & van Koppen, M. (2002), Pronominal doubling and the structure of the left periphery in southern Dutch, in S. Barbiers, L. Cornips & S. v. d. Kleij, eds, 'Syntactic microvariation', <http://www.meertens.knaw.nl/books/synmic/>.

- van Craenenbroeck, J. & van Koppen, M. (2008), Pronominal doubling in Dutch dialects: big DPs and coordinations, in S. Barbiers, O. Koenenman, M. Lekakou & M. van der Ham, eds, 'Microvariation in syntactic doubling.', Vol. 36 of *Syntax and Semantics*, Emerald, Bingley, pp. 207–249.
- van Craenenbroeck, J. & van Koppen, M. (2021), Microvariation and parameter hierarchies. Ms. KU Leuven & Meertens Institute & Utrecht University.
- van Craenenbroeck, J., van Koppen, M. & van den Bosch, A. (2019), 'A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters', *Language* **95**, 333–370.
- van Koppen, M. (2017), Complementizer agreement, in M. Everaert & H. van Riemsdijk, eds, 'The Wiley-Blackwell Companion to Syntax—Second Edition', Wiley-Blackwell, pp. 923–962.
- Zanuttini, R., Wood, J., Zentz, J. & Horn, L. (2018), 'The Yale Grammatical Diversity Project: Morphosyntactic variation in North American English', *Linguistics Vanguard* **4**(1).
- Zwart, J.-W. (1992), Dutch expletives and small clause predicate raising, in K. Broderick, ed., 'Proceedings of North East Linguistic Society 22', GLSA, Amherst, MA, pp. 477–491.