Chapter 4

Quantitative approaches to syntactic variation

Jeroen van Craenenbroeck[1,2] and Marjo van Koppen[2,3]

[2] Meertens Institute
[1] KU Leuven/CRISSP
[3] Utrecht University/ILS

Abstract

This chapter discusses quantitative approaches to studying syntactic variation, more specifically approaches that pursue a combined quantitative-qualitative methodology, integrating components both from the formal-theoretical and the computational-statistical tradition. We first introduce a number of case studies of such integration, before zooming out and highlighting the advantages and benefits offered by a combined quantitative-qualitative approach and listing some of the fundamental theoretical issues it raises for the study of variation.

4.1 Introduction

This chapter focuses on quantitative approaches to syntactic variation. We discuss the advantages of adopting a combined quantitative-qualitative methodology towards investigating variation, and offer a number of concrete examples of such integration. Before we do so, however, we need to delineate our subject matter and make clear what this chapter is *not* about. While the phrase 'quantitative approaches to syntactic variation' immediately excludes from the discussion certain types of research—e.g. we will not address the (extensive) literature analyzing phonological variation from a quantitative point of view—it is still sufficiently general so as to cover more (types of) research than can reasonably be fit into one handbook chapter. As a result, we first need to circumscribe the topic of this chapter more precisely. Two subdisciplines of linguistics we will systematically leave undiscussed are diachronic syntax and (first and second) language acquisition. While the use of quantitative methods is prevalent in these domains, they are already well-described and well-documented (see for example Ledgeway and Roberts (2017) and De Villiers and Roeper (2011)), and both diachrony and language acquisition have their own separate chapter in this handbook (see Chapters 8 and 9 respectively).

Our main focus in this chapter will be on studies that combine a quantitative-statistical approach on the one hand with a qualitative formal analysis on the other. Put differently, approaches whereby visible, countable properties of language are analyzed in order to gain a deeper understanding of the abstract representations and derivations that underlie those surface properties. This implies that purely descriptive work—even when it makes extensive use of quantitative methods—

will not feature prominently in this chapter (see for example Spruit (2008) or Jeszenszky et al. (2017)), and nor will usage-based approaches to syntactic variation: while quantitative methodology and corpus-based data collection form a cornerstone of this type of research, the existence of an abstract grammatical knowledge system that is independent from actual language use or performance is called into question in these approaches, thus rendering moot the issue of the interaction between the two components. Some representative work in this tradition includes Szmrecsanyi and Kortmann (2009), Szmrecsanyi (2013), and De Troij et al. (2023).

The reason for focusing specifically on the interaction between quantitative and qualitative approaches is that we believe this to be an especially promising area of research, one that shows great potential for deepening our understanding of syntactic variation in future research (see also Cornips and Corrigan (2005)). The chapter is organized as follows. In the next section we introduce and discuss a number of concrete examples of integrated quantitative-qualitative analyses, organized according to the degree of integration between the two, while section 4.3 presents some more general considerations on the topic: we discuss the advantages of a combined approach and highlight some of the fundamental theoretical issues it raises. Section 4.4 concludes.

4.2 Degrees of integration

In this section we discuss a number of concrete examples of integrated quantitative-qualitative approaches to syntactic variation. We focus on the main results and on the techniques used to reach those results. We have organized the section according to the increasing degree of integration between the qualitative and the quantitative component: in subsection 4.2.1 we focus on approaches whereby statistical means are used to detect (typically: geographical) correlations in the data, which then serve as the basis for a theoretical account. Subsection 4.2.2 focuses on case studies whereby aspects of a theoretical analysis are used as independent variables in a predictive statistical model. Finally, in subsection 4.2.3 we introduce analyses that are concerned with model comparison and model selection, i.e. whereby entire formal-syntactic analyses can be compared and evaluated against one another or against a baseline.

4.2.1 Correlations

The recurring theme running through the case studies discussed in this subsection is the idea that correlations in the data between two or more properties suggest a common theoretical source for those properties. In so doing, this approach relies on the traditional notion of a parameter as a single choice point in the grammatical system that simultaneously manifests itself in more than one surface phenomenon (see Rizzi (1986) for the pro-drop parameter as an early and famous example of this line of thinking). What these accounts moreover have in common, and the reason why they are included in this chapter, is that these correlations are calculated in an automated way over large data sets, either by directly calculating the correlation coefficients between pairs of properties, or by using exploratory statistical techniques such as multidimensional scaling, correspondence analysis, or hierarchical clustering.

A first example of this approach is Wood and Zanuttini (2018). They focus on the non-standard use of datives in dialects of American English, a phenomenon mainly attested in the south east of the United States. An example is given in (1), where the dative pronoun *you* occurs in a presentative sentence.

(1) Here's you a piece of pizza.

Wood and Zanuttini's data are based on questionnaires, where informants (recruited via Mechanical Turk) were asked to rate sentences on a scale of 1 to 5. These data are then visualized in a three-step procedure: first, the clearly negative (1 or 2 on a scale of 1 to 5) and clearly positive (4 or 5) answers are directly plotted on a geographical map. Second, the areas in between the questionnaire points are color-coded on the basis of data interpolation, a technique to estimate missing values in a data set based on other, known values. Third, the questionnaire data are analysed using a technique from geostatistical analysis sometimes referred to as hot spot analysis, in order to determine which of the attested geographical patterns are statistically reliable. In combination, these three techniques yield maps such as the one illustrated in Figure 4.1.

This map represents the results of one question from Wood and Zanuttini's questionnaire, namely the informants' rating of the sentence in (1). The green dots on the map represent judgments of 4 or 5, while the black dots indicate low scores (1 or 2). The shade of blue is a measure of the interpolated data: the darker the shade, the higher the estimated rat-

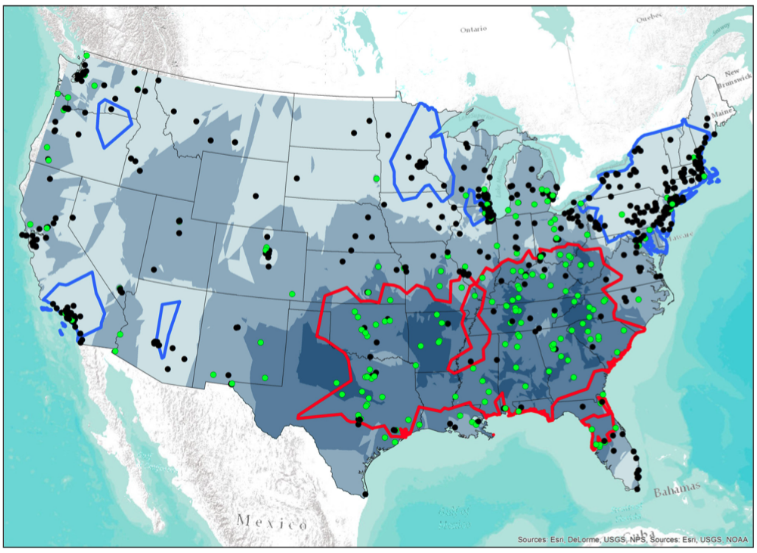

Ontario
Quebec
New Brunswick
Maine
Mexico
Bahamas
Sources: Esri, DeLorme, USGS, NPS, Sources: Esri, USGS, NOAA

Figure 4.1 Geographic distribution of *Here's you a piece of pizza* (from Wood and Zanuttini (2018:8))

ing. Finally, the zone demarcated in red is a hot spot: an area where the number of high scores is greater than would be expected by chance. By contrast, in the areas delineated with a blue line, the number of low scores is statistically significant. This three-fold technique of data analysis and data visualization ensures that Wood and Zanuttini's discussion of non-standard dative constructions is on a firm empirical footing, and they then proceed to use these visualizations as arguments on which to base part of their analysis. More specifically, consider the three maps in Figure 4.2.

All three of these maps represent constructions that (a) are typical of the south of the US, and (b) involve dative expressions. Relevant examples are given in (2).

(2) a. He has him a new car.
b. Here's you a piece of pizza.
c. We are looking for him a new home.

Wood and Zanuttini refer to these constructions as Personal Datives (2a), Southern Dative Presentatives (2b), and Extended Benefactives (2c) respectively. The maps in Figure 4.2 clearly show that the latter

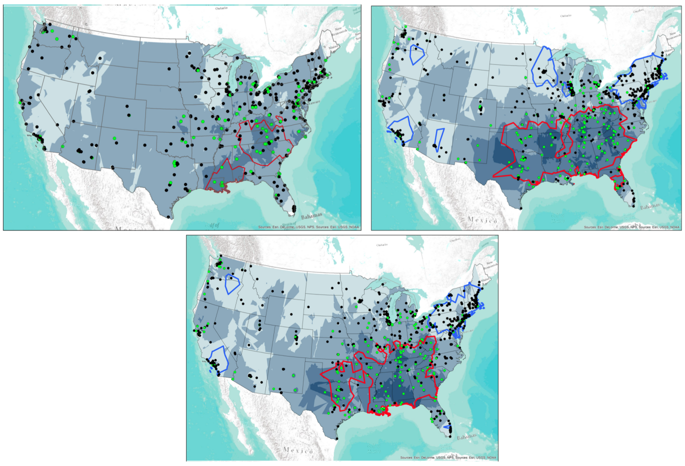


Figure 4.2 Geographic distribution of *He has him a new car* (top left), *Here's you a piece of pizza* (top right), and *We are looking for him a new home* (bottom) (from Wood and Zanuttini (2018:4,8,13))

two are much more similar in their geographical distribution than Personal Datives are to either of them. This observation informs Wood and Zanuttini's formal analysis of these data: they identify two points of variation in their account, one that sets Personal Datives apart from standard English, and one that groups together Southern Dative Presentatives and Extended Benefactives. Both are related to the low applicative phrase (ApplP): in order for Personal Datives to be allowed, the Appl-head has to license the presence of a ϕP in its specifier, while Southern Dative Presentatives and Extended Benefactives require that the entire low ApplP has the same distribution as a DP.[1]

As Wood and Zanuttini are themselves well aware (Wood and Zanuttini 2018:12), the line of reasoning just sketched risks falling prey to the old adage that correlation is not causation: just because two properties have the same geographical distribution, that does not mean that they are theoretically related as well—especially when the phenomena in question occur in a coherent contiguous geographical region. This holds all the more so for microvariation, where the varieties under consideration are genealogically closely related and the geographic distances between them are very small.[2] The next accounts we discuss alleviate

[1] See the original paper for further technical details.

[2] See in this respect Haspelmath (2008:86n8), who points out that by studying genealogically related languages one "runs the risk [...] [to] discover shared

that worry to a certain degree, in that geographic information is used in the analysis as a purely categorical variable—i.e. the answer to the question: is phenomenon X attested in location Y or not?—irrespective of the geospatial patterns this information represents. This means two phenomena will pattern together in the quantitative analysis also if the locations they both occur in are very disparate and do not form a contiguous area, i.e. when an alternative account in terms of language contact or shared history seems less likely.

A first example of this approach is Iosad and Lamb (2020). They present a dialectometric analysis of Scottish Gaelic morphology and show how such a quantitative analysis can lead to new insights into the grammatical system of the language. The data set consists of 55 morphological properties from 201 different locations extracted from the Linguistic Survey of Scotland (Bosch 2006), to which they apply a correlation analysis: the correlation is calculated between each pair and the significance of this correlation is tested. The outcome of this analysis is represented in the plot in Figure 4.3, where the (dark) red hues indicate a high degree of correlation (significant at the $p < 0.1$-level). Note that in this correlation matrix, unlike in the approach of Wood and Zanuttini, the actual geographical distribution plays only a secondary role: two properties correlate highly when they are attested in—and absent from—the same dialect locations, irrespective of whether those dialect locations are close to one another or geographically far apart.

Iosad and Lamb then use these correlations to inform their formal-theoretical analysis of these data. The Scottish Gaelic dialects are characterized by a high degree of morphological variability, which is often described in terms of attrition or even language death. From a theoretical point of view, however, such changes can have different types of causes. Consider in (3) a schematic spell-out rule of the type used in Distributed Morphology (Halle and Marantz 1993), the theoretical framework adopted by Iosad and Lamb.[3]

(3) $[\mathrm{F}_1] \Leftrightarrow /\mathrm{E}_1/$

What this rule states is that a particular grammatical feature F_1 (e.g. [+PLURAL]) is realized as a specific exponent E_1 (e.g. the suffix *-s*). Now assume that in some varieties of Scottish Gaelic E_1 is missing. At a

innovations that have purely historical explanations, rather than properties that are shared because of the same parameter setting". See also Pescarini (2019), discussed below.

[3] In DM-parlance, the rule in (3) is known as a Vocabulary Item.

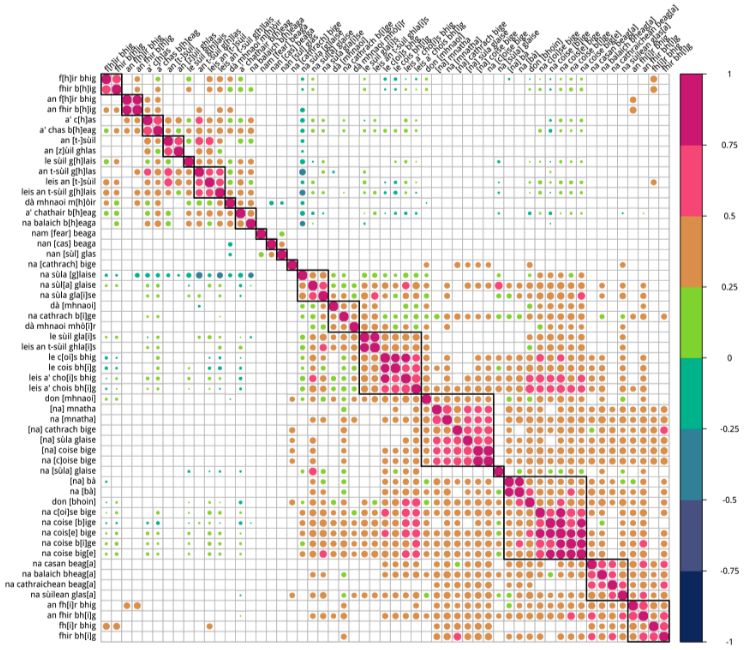


Figure 4.3 Correlation plot of features (from Iosad and Lamb (2020:22, Figure 8))

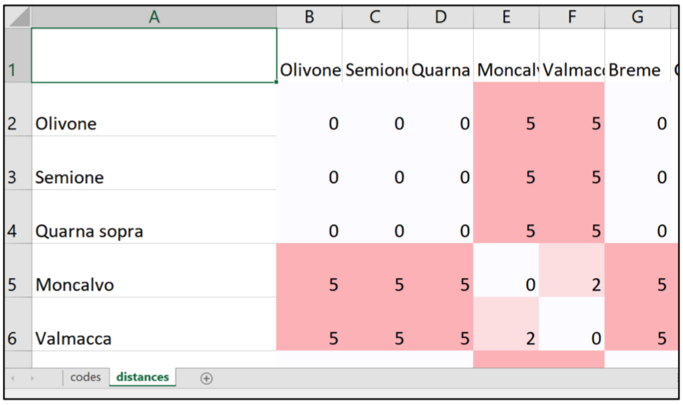
descriptive level this might be characterized as a case of attrition, but theoretically, this absence could have various causes. For one, it could be that the feature F_1 is missing from the grammar and as a result, the rule in (3) can no longer apply. Alternatively, however, the change might affect the rule itself: maybe E_1 has been replaced by a different (possibly null) exponent. Iosad and Lamb show that the correlations illustrated in Figure 4.3 can help decide between these two options. For example, the expression of vocative as lenition in singular masculine nouns of the first declension is strongly correlated with the expression of vocative as lenition in masculine singular adjectives. This suggests that whatever mechanism is responsible for this variation, it should be sufficiently general to cover both nouns and adjectives. At the same time, in dialects where vocative-as-lenition is absent, we cannot simply claim that the feature [vocative] is missing from the grammar, because lenition does not correlate particularly strongly with slenderisation, another exponent of vocative in Scottish Gaelic. As a result, Iosad and Lamb conclude that in this case it is only the exponence rule—particularly the right-hand part of that rule, cf. (3)—that is affected by the dialect change, not the underlying grammatical system. The opposite conclusion holds for variation in the expression of feminine gender. Here, it is not only lenition on nouns and adjectives that are correlated, but also *t*-sandhi after the definite article. This suggests that this instance of language change does not affect specific mechanisms of exponence, but rather the underlying grammatical category, in this case the feminine gender feature: with this feature gone, the exponence rule no longer applies, and all possible exponents are lost simultaneously. More generally, this paper provides a good illustration of the type of research highlighted in this subsection: quantitatively established correlations in the data set are seen as informative about the underlying grammatical system.

We find the same line of reasoning in Van Craenenbroeck and van Koppen (2021). They start from a data set involving 10 dialect phenomena in 267 dialect locations in Belgium, the Netherlands, and the north of France[4] and they use correspondence analysis to reduce the dimensionality of this data set to a three-dimensional representation of those ten phenomena. When two phenomena are close to one another in this 3D-space, they have a highly comparable geographical distribution. Van Craenenbroeck & Van Koppen then interpret each of the three dimensions of this space in formal-theoretical terms, by identifying the parameters responsible for deriving the variation in each dimension. Once again, then, quantitative correlations in the data drive a formal-theoretical analysis that is built based on these correlations.

Pescarini's (2019) account of subject clitics in northern Italian dialects is highly similar in spirit, but very different in execution. Pescarini first encodes the subject clitic paradigm of each dialect, in particular in terms of the gaps and syncretism patterns it contains, into an abstract code. For example, a dialect encoded as G023456S45 has a gap (G) in the first person singular—in other words, there is no first person singular subject clitic—and a syncretism (S) in the first and second plural—represented by the numbers 4 and 5 in the code. The degree of similarity between two dialects can now be calculated by looking at the edit distance—also known as the Levenshtein distance—between their encodings. The edit distance is a measure of the number of steps or operations needed to get from one expression to another. For example, to go from G023456S45

[4] These data stem from the Syntactic Atlas of Dutch Dialects (SAND), cf. Barbiers et al. (2005), Barbiers *et al.* (2006), Barbiers et al. (2008).

to G120456S, four steps are needed: two substitutions (from 0 to 1, and from 3 to 0) and two deletions (the 4 and 5 at the end). On the basis of these comparisons Pescarini creates a distance matrix, a small portion of which is shown in Figure 4.4.



	A	B	C	D	E	F	G
1		Olivone	Semion	Quarna	Moncal	Valmac	Breme
2	Olivone	0	0	0	5	5	0
3	Semione	0	0	0	5	5	0
4	Quarna sopra	0	0	0	5	5	0
5	Moncalvo	5	5	5	0	2	5
6	Valmacca	5	5	5	2	0	5

Figure 4.4 Part of the distance matrix of norther Italian dialects based on their clitic inventory (from Pescarini (2019:273))

In this table, the dialects under investigation constitute both the row and the column labels. The values in the cells represent the edit distance between that pair of dialects.[5] Pescarini then uses a Mantel test to measure the correlation between this distance matrix and one whereby the cells contain the actual geographical distances between the dialect locations. In other words, he measures the correlation between the linguistic distance and the geographical distance. Given that this correlation is very low—the result of the Mantel statistic is an index of 0.05931—Pescarini concludes that the variation found in subject clitic systems in northern Italian dialects "cannot be accounted for under a pure geolinguistic explanation" (Pescarini 2019:274).

At the same time, the variation in subject clitic syncretism does not appear to be random, in that Pescarini's study reveals clear empirical generalizations. For example, by far the most common syncretism is between first person singular and first person plural: it occurs in 80% of the dialects. Moreover, a large majority of those dialects—almost 84%—also include the second person plural in that syncretism. At the other

[5] Note that the diagonal contains only zeroes because the edit distance from a dialect to itself is 0, and that the table is symmetric across the diagonal because the distance from A to B is identical to the distance from B to A.

end of the spectrum, the second person singular is hardly ever syncretic with other persons, and in the rare cases that it is, the syncretism always involves multiple other persons. Given that geographical proximity seems unable to account for these patterns, Pescarini argues that understanding them requires a more abstract account in terms of theoretical constraints. In so doing, he inverts the perspective discussed above: it is not the case that geography-based correlations in the data are assumed to be informative about the underlying formal analysis. Rather, it is the *lack* of a coherent geographic signal in the data that is seen as indicative that a formal account is needed to analyze the empirical patterns that have emerged. Pescarini's line of reasoning comes out quite clearly in the following quote:

> "Historically, the basin of the river Po and the surrounding mountains have always been a well-interconnected area, where people and goods circulated rather freely despite the geopolitical fragmentation. Given this socio-historical scenario, one would expect linguistic innovations to spread homogeneously in contiguous areas regardless of biolinguistic constraints on the make-up of pronominal inventories. Alternatively, one may hypothesize that patterns of gaps and syncretism [...] are due to a biolinguistic constraint preventing or hindering the externalization of certain clitic forms. Then one would expect to find the same pattern scattered in non-contiguous dialects (Poletto's 2003 leopard spots), regardless of socio-historical factors." (Pescarini 2019:271)

In a way, then, we have come full circle: while Wood and Zanuttini's (2018) approach is mostly based on coherent and clearly identifiable geographical regions, the correlation analyses of Iosad and Lamb (2020) and Van Craenenbroeck and van Koppen (2021) leave room for a less direct effect of geography, while Pescarini (2019) takes it one step further still and suggests theoretical syntacticians should primarily focus on linguistic phenomena that do *not* have a clear geographical pattern. Only then can we be reasonably confident that the observed patterns are not due to grammar-external properties like language contact or a shared history. What all of these accounts still have in common, though, is that the qualitative and the quantitative part of the analysis are juxtaposed rather than truly integrated: first the data are quantitatively analysed, and then a qualitative account is proposed that integrates some of the findings of that first step. In the case studies discussed in the next two subsections, the integration between the two approaches is more direct.

4.2.2 Theory as predictors

What the accounts in this subsection have in common is that aspects of the formal-theoretical analysis are directly integrated into the quantitative analysis. In terms of the techniques and methodologies used, this typically involves the transition from descriptive and exploratory statistical methods—as discussed in the previous subsection—to inferential and predictive statistics. The most common—but by no means the only—implementation of this scenario is a regression analysis whereby quantifiable parameters of formal-theoretical concepts are used as independent variables, typically alongside geographical and/or social variables. A good example of this approach is Burnett et al. (2018). They focus on variation between negative quantifiers (4a) and negative polarity items (4b) in object position.

(4) a. They have no friends.
b. They don't have any friends.

The traditional account of this alternation is that it presents an intermediate stage in the historical development from the older pattern in (4a) to the newer one in (4b), a development moreover that is conditioned by frequency: more frequent verbs are more conservative and show higher counts of negative quantifiers, while less frequent verbs typically co-occur with negation and a negative polarity item (Tottie 1991a,b). As pointed out by Burnett et al., however, this frequency-based account faces a number of challenges. For example, in some corpora the rates of negative quantifiers versus negative polarity items does not correlate with the frequency of the accompanying verbs, and in French—where the opposite process seems to be taking place, i.e. from negative polarity items to negative quantifiers—the change seems to be driven by the *more* frequent items rather than by the less frequent ones. What Burnett et al. propose instead, is that the choice between (4a) and (4b) is subject to more abstract, structural conditions. Following the lead of Kayne (1998), they suggest that English is like mainland Scandinavian in that negative quantifiers undergo object shift out of the VP, while negative polarity items remain VP-internal. They then encode the data—a set of 1154 utterances from the speech of 88 speakers extracted from the Toronto English Archive—according to this abstract parameter, based on the type of verb or construction the negative element co-occurs with: negative objects that are embedded under more than one verb (5a), that occur in a non-finite sentence (5b), or that are embedded under another element such as a preposition (5c), are all coded as being inside the VP, whereas negative objects that are not embedded in that way—as in (6)—are coded as being higher than the VP.

(5) a. I don't envy any of them.
b. They were worried there were going to be no French Catholics left.
c. We're under no obligation.

(6) a. There were no jobs to be had.
b. It was nothing like that.

Burnett et al. then proceed to formulate the following prediction:

(7) **Prediction of (Soft) Negative Object Shift Analysis**
We should find a significantly higher rate of Neg-Qs in utterances that could be parsed as having the negative indefinite appear higher than the VP than in those utterances in which the indefinite clearly remains within the VP.

This prediction is clearly confirmed by Burnett et al.'s experiment: they construct a mixed-effect regression model with syntactic position as one of the predictor variables, and it emerges as the most significant factor. This in turn leads Burnett et al. to a novel interpretation of the historical change underlying the constructions in (4): rather than a wholesale replacement operation of negative quantifiers by negative polarity items, the change only affects VP-internal elements, and the use of negative quantifiers VP-externally is not affected. In other words, the grammatical system is converging towards a stable state whereby negative quantifiers and negative polarity items exist side by side in object position. Crucially, though, this is a conclusion that could only be reached thanks to the combined quantitative-qualitative approach adopted by Burnett et al.

A second illustration of this line of research is Van Craenenbroeck et al. (2019). Their empirical focus is word order in clause-final verb clusters in dialects of Dutch. As is well-known, verbs—finite and nonfinite alike—cluster at the end of the (embedded) clause in Dutch. An example of a three-verb cluster is given in (8).

(8) Ik vind dat iedereen **moet kunnen zwemmen**.
I find that everyone must can swim
'I think everyone should be able to swim.'

In principle, three verbs can be ordered in six (three factorial) different ways, but in practice, the particular type of verb cluster illustrated in (8)—two modal auxiliaries and a lexical main verb—only shows up in four of those six orders in the dialects of Dutch:

(9) a. Ik vind dat iedereen moet kunnen zwemmen.
b. Ik vind dat iedereen moet zwemmen kunnen.
c. Ik vind dat iedereen zwemmen moet kunnen.
d. Ik vind dat iedereen zwemmen kunnen moet.
e. *Ik vind dat iedereen kunnen zwemmen moet.
f. *Ik vind dat iedereen kunnen moet zwemmen.

Moreover, not every dialect allows all four of these orders: some allow only one, others two or three, and not always the same one or two or three. More generally, Van Craenenbroeck et al. examine six different cluster types for a total of 31 different cluster orders in 267 dialects of Dutch—data taken from the SAND-atlas, Barbiers *et al.* (2006)—and they find substantial word order variation in dialect Dutch verb clusters. The first step of their quantitative analysis is very much in line with the accounts described in the previous subsection: they use exploratory statistical techniques—Correspondence Analysis to be precise—to reduce the dimensionality of the data set and identify the main tendencies and correlations. This yields the plot illustrated in Figure 4.5.

Each of the 31 cluster orders is represented in this plot. The one in (9d) for example can be found at the left edge of the plot, close to the x-axis. When two cluster orders are close together in this plot, they typically have the same distribution—irrespective of whether the locations they occur in form a contiguous geographical region, see the discussion in subsection 4.2.1 above—and when they are far apart they tend not to occur in the same dialect locations. Rather than directly build a formal analysis on these observations, however, Van Craenenbroeck et al. integrate into the quantitative analysis principles and mechanisms extracted from the theoretical literature on verb cluster ordering. For example, some analyses derive word order variation through leftward VP-movement starting from a head-initial base (Barbiers 2005), whereas others assume VPs are head-final and non-head final orders are derived via rightward head movement (Evers 1975). In total, Van Craenenbroeck et al. integrate 64 linguistic variables from 11 analyses of verb clusters into their analysis. The degree to which a specific linguistic mechanism or principle aligns with the patterns found in the data set can now be visually represented

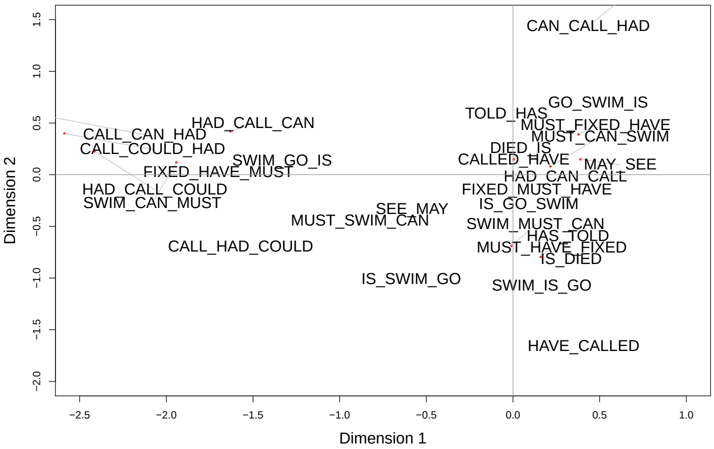


Figure 4.5 Two-dimensional representation of the SAND verb cluster data using Correspondence Analysis (from Van Craenenbroeck et al. (2019:347))

by color-coding the plot in Figure 4.5 according to that principle. An example is given in Figure 4.6.

This figure contains the same plot as Figure 4.5, but color-coded according to one of the ingredients of Haegeman and van Riemsdijk (1986)'s formal analysis of word order variation in verb clusters. They propose that one of the parameters regulating such variation concerns the relationship between modal verbs and their complement: in some dialects the two undergo inversion, whereas in others they do not. If so, we would expect cluster orders that involve modal inversion to pattern together and differently from cluster orders that do not involve modal inversion. This is confirmed by the color-coding in Figure 4.6: the distinction between red and green orders—the black ones do not contain a modal and hence are irrelevant for this criterion—is strongly correlated with the first dimension of the Correspondence Analysis. This visual result can be further corroborated by a more precise numeric one: Van Craenenbroeck et al. calculate, for each combination of linguistic variable and CA-dimension, the squared correlation ration (η^2), a measure for the proportion of variance on that particular dimension that is explained by that linguistic variable. Haegeman and van Riemsdijk's (1986) modal inversion parameter has an η^2 of 0.599 in the first CA-

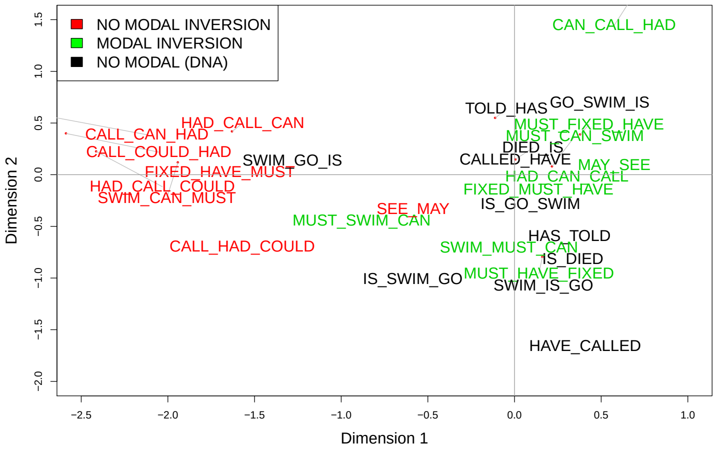


Figure 4.6 Two-dimensional representation of the SAND verb cluster data color-coded according to Haegeman and van Riemsdijk's (1986) modal inversion parameter

dimension, which is the fourth highest among the 64 linguistic variables. Abstracting away from this specific result, however, it should be clear by now that—and how—Van Craenenbroeck et al. directly integrate aspects of formal, qualitative analyses into their quantitative analysis, just like Burnett et al. (2018) did.

A similar approach is adopted by Pescarini (2022). He uses logistic regression to examine negative marking in central Romance dialects. The independent variables contain both geographical and grammatical information, and while the former systematically comes out as significant, models that include both types of information typically outperform models containing only geographical factors. Other work in the same general vein includes Samo and Merlo (2019, 2021), who examine intervention effects in cleft and relative clauses in French, Italian, and English. Just like Burnett et al. (2018), Samo and Merlo base their research on (syntactically annotated) corpus data rather than questionnaire results: they show that differences in corpus frequency between object and subject clefts/relative clauses can be explained as a result of a Rizzian intervention effect affecting the former—where one argument has to move across another—but not the latter.

Finally, an approach that also deserves mention in this subsection is the co-called Parametric Comparison Method (PCM) developed by Giuseppe Longobardi and collaborators (see for example Longobardi (2003), Longobardi and Guardiano (2009), Guardinao and Longobardi (2005), Guardiano et al. (2016, 2020)). The method starts out from the idea that theory-based comparative research should be based on "studying relatively *many* parameters across relatively *many* languages within a *single* module of grammar" (Longobardi (2018:522), emphasis in the original). That one should look at a sufficiently large number of languages and parameters in order to gain a certain degree of coverage and representativeness speaks to reason, but the choice to limit those investigations to a particular empirical domain is new, and it allows one to focus on a particular aspect of parameters that typically receives less attention, namely their interdependence. The primary empirical domain of choice in the PCM-literature is the DP, with the famous 'Table A' as one of its most tangible instantiations. A small portion of that table is given in Figure 4.7.

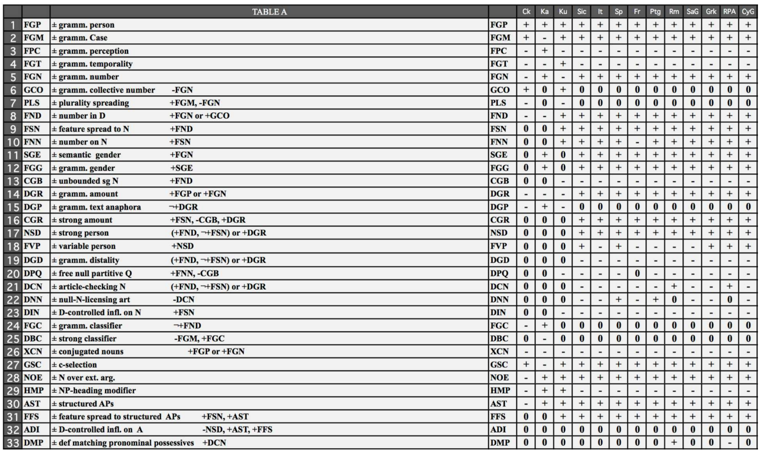


		TABLE A			Ck	Ka	Ku	Sic	It	Sp	Fr	Ptg	Rm	SaG	Grk	RPA	CyG
1	FGP	± gramm. person		FGP	+	+	+	+	+	+	+	+	+	+	+	+	+
2	FGM	± gramm. Case		FGM	+	-	+	+	+	+	+	+	+	+	+	+	+
3	FPC	± gramm. perception		FPC	-	+	-	-	-	-	-	-	-	-	-	-	-
4	FGT	± gramm. temporality		FGT	-	-	+	-	-	-	-	-	-	-	-	-	-
5	FGN	± gramm. number		FGN	-	+	-	+	+	+	+	+	+	+	+	+	+
6	GCO	± gramm. collective number	-FGN	GCO	+	0	+	0	0	0	0	0	0	0	0	0	0
7	PLS	± plurality spreading	+FGM, -FGN	PLS	-	0	-	0	0	0	0	0	0	0	0	0	0
8	FND	± number in D	+FGN or +GCO	FND	-	-	+	+	+	+	+	+	+	+	+	+	+
9	FSN	± feature spread to N	+FND	FSN	0	0	+	+	+	+	+	+	+	+	+	+	+
10	FNN	± number on N	+FSN	FNN	0	0	+	+	+	+	-	+	+	+	+	+	+
11	SGE	± semantic gender	+FGN	SGE	0	+	0	+	+	+	+	+	+	+	+	+	+
12	FGG	± gramm. gender	+SGE	FGG	0	+	0	+	+	+	+	+	+	+	+	+	+
13	CGB	± unbounded sg N	+FND	CGB	0	0	-	-	-	-	-	-	-	-	-	-	-
14	DGR	± gramm. amount	+FGP or +FGN	DGR	-	-	-	+	+	+	+	+	+	+	+	+	+
15	DGP	± gramm. text anaphora	¬+DGR	DGP	-	+	-	0	0	0	0	0	0	0	0	0	0
16	CGR	± strong amount	+FSN, -CGB, +DGR	CGR	0	0	0	+	+	+	+	+	+	+	+	+	+
17	NSD	± strong person	(+FND, ¬+FSN) or +DGR	NSD	0	0	0	+	+	+	+	+	+	+	+	+	+
18	FVP	± variable person	+NSD	FVP	0	0	0	+	-	+	-	-	-	-	+	+	+
19	DGD	± gramm. distality	(+FND, ¬+FSN) or +DGR	DGD	0	0	0	-	-	-	-	-	-	-	-	-	-
20	DPQ	± free null partitive Q	+FNN, -CGB	DPQ	0	0	-	-	-	-	0	-	-	-	-	-	-
21	DCN	± article-checking N	(+FND, ¬+FSN) or +DGR	DCN	0	0	0	-	-	-	-	-	+	-	-	+	-
22	DNN	± null-N-licensing art	-DCN	DNN	0	0	0	-	-	+	-	+	0	-	-	0	-
23	DIN	± D-controlled infl. on N	+FSN	DIN	0	0	-	-	-	-	-	-	-	-	-	-	-
24	FGC	± gramm. classifier	¬+FND	FGC	-	+	0	0	0	0	0	0	0	0	0	0	0
25	DBC	± strong classifier	-FGM, +FGC	DBC	0	-	0	0	0	0	0	0	0	0	0	0	0
26	XCN	± conjugated nouns	+FGP or +FGN	XCN	-	-	-	-	-	-	-	-	-	-	-	-	-
27	GSC	± c-selection		GSC	+	-	+	+	+	+	+	+	+	+	+	+	+
28	NOE	± N over ext. arg.		NOE	-	+	+	+	+	+	+	+	+	+	+	+	+
29	HMP	± NP-heading modifier		HMP	-	+	+	-	-	-	-	-	-	-	-	-	-
30	AST	± structured APs		AST	-	+	+	+	+	+	+	+	+	+	+	+	+
31	FFS	± feature spread to structured APs	+FSN, +AST	FFS	0	0	+	+	+	+	+	+	+	+	+	+	+
32	ADI	± D-controlled infl. on A	-NSD, +AST, +FFS	ADI	0	0	0	0	0	0	0	0	0	0	0	0	0
33	DMP	± def matching pronominal possessives	+DCN	DMP	0	0	0	0	0	0	0	0	+	0	0	-	0

Figure 4.7 Top left-hand portion of Table A (from Longobardi (2018:523))

The rows in this table are parameters that are relevant to the nominal domain—such as whether or not [person] is grammaticalized in the DP—and the columns represent languages. Each parameter can be set to +, −, or 0, in which case its value is automatically set or determined as a result of another parameter setting, i.e. when there is an interdepence between two parameters. The original goal of the PCM was to determine to what extent a theory-based syntactic classification of languages can be used to reconstruct their genealogical dependencies. It has turned out to be very successful in this respect: a PCM-based classification analyzed with computational methods of phylogenetic analysis and taxonomic representation is able to capture over 85% of a golden standard of a typology of language relatedness based on lexical (cognacy) characteristics. At the same time, however, the results from the PCM can also feed back into the theoretical literature on syntactic variation. For example, it has led to a rethinking of parameter theory itself, with Longobardi (2018) arguing for the abandonment of parameters as a list of binary choice points and replacing that view with one of 'parameter schemata': a limited list of templates or types that parameters can fall into, irrespective of the specific feature or empirical domain targeted by that parameter. Similarly, Kazakov et al. (2017) apply machine learning methods to parameter tables such as the one in Figure 4.7 to reveal previously unknown dependencies between parameters, thus identifying redundant parameters and reducing the search space for the language-learning child.

4.2.3 Model comparison and model selection

The accounts discussed in this third and final subsection show the deepest level of integration between quantitative and qualitative approaches to syntactic variation. Although they are quite heterogeneous in nature, there is one central theme that connects them and that is the fact that they can be used to (dis)confirm, compare, and select entire analyses, as opposed to individual ingredients or components of those analyses. In terms of the methods used, the proposals discussed in this section continue to make use of regression models, but classifiers such as Bayesian algorithms or k-nearest neighbors classification are used as well, among a host of other, more specific tools.

A prototypical example is Merlo's (2015) discussion of (possible analyses of) Greenberg's (1963) Universal 20. More generally, the work of Paola Merlo and her collaborators deserves special mention in a chapter devoted to theoretically inspired quantitative approaches to variation, as they have been pioneering this approach for almost ten years now. Merlo terms the enterprise 'quantitative computational syntax' (Merlo 2016), a moniker that nicely characterizes the approach as the combination of (theoretical) syntax with quantitative computational methodologies. Merlo (2015) focuses on Greenberg's (1963) Universal 20:

(10) **Greenberg's Universal 20**
When any or all the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in this order. If they follow, the order is exactly the same or its exact opposite.

As is well-known, while Greenberg got the basics of DP-internal word order right—the orders he identifies are indeed attested in the world's languages and there is more word order freedom postnominally than prenominally—there is much more word order variation in this domain than is suggested by (10). Based on updated counts and more recent and extensive typological information (Cinque 2005, Dryer 2006), Merlo sets out to compare three accounts of the variation attested in this domain: Cinque (2005), Cysouw (2010), and Dryer (2006) (later published as Dryer (2018)). The central question is how well these accounts predict the typological distribution—i.e. the frequency—of all the possible permutations of the sequence Dem-Num-A-N. The experiment proceeds in four steps (Merlo 2015:326):

(11) a. Formalise the properties and operations posited by a model of word order as simple primitive features with a set of associated values;
b. Encode each word order as a vector of instantiated primitives defined by the model;
c. Learn the model through a learning algorithm on a subset of the data;
d. Run the model on previously unseen data to test generalisation ability.

In other words, first the analysis of each word order is encoded as a vector for each model, the classifier is trained on a subset of those vectors, and then the question is whether based on that information, it can correctly predict the frequency of unseen word orders, i.e. of word orders that were not part of the training data. In order to make this more concrete, consider how Cinque (2005) would derive the order N-Num-A-Dem: starting from a universal base order Dem>Num>A>N—where '>' represents c-command—this order first requires NP-movement to the left of A and then to the left of Num, followed by N+Num+A-movement to the left of Dem. Merlo encodes this analysis as in (12).

(12) AN NumA DNum np whose-pp R

The first three values in this vector represent the three (external) merge operations needed to derive this order: A merges with N, Num with (the phrase headed by) A, and D(em) with (the phrase headed by) Num.[6] The fourth value characterizes the partial movement operations needed to derive this order, whereby 'partial' is interpreted as 'does not target a position to the left of Dem'. As outlined above, the derivation of this order in Cinque's analysis requires NP-movement to the left of A and Num, but not all the way to the left of Dem, so the analysis contains partial NP-movement. The fifth value in the vector lists complete movement operations, i.e. ones that do reach the left edge of the DP. In this case the movement of N+Num+A to the left of Dem is of this type. Assuming (as does Cinque) that all movement operations are triggered by N, this one involves pied-piping of the *whose picture*-type, hence the encoding as 'whose-pp'. Finally, the 'R' at the end of the vector is not part of the encoding of Cinque's analysis, but represents the dependent variable: it indicates that this particular order is typologically rare.

Merlo uses the same encoding mechanism for the remaining 23 word orders, and creates a similar vector representation of the analyses of Cysouw (2010) and Dryer (2006). She then uses a Naive Bayes classifier with ten-fold cross-validation to test the performance of these three models. A classification task involves predicting which class a particular instance belongs to based on the properties it has. The classifier is first trained on a set of known instances, i.e. cases where both the properties and the correct class are given. Based on that training phase, it is given new, unseen instances, which it is asked to classify. The number of correct classifications is a measure for the success of the model. A Naive Bayes classifier is based on Bayes' theorem, and its most noticeable feature is that it assumes all the properties under discussion are independent of one another—more on that below. For example, in Cinque's analysis, the merge order is independent of whether or not partial movement takes place or what type of movement this is. Cross-validation is a training and testing protocol: it implies that the data is randomly divided into a number of subsets (ten in this case), and that the experiment is run multiple times, with each subset in turn serving as the unseen testing data. Merlo runs the experiment multiple times: once at the type level (where the models have to predict the frequency class of word orders) and once

[6] In Cinque's analysis, this is the only possible merge order, but in order to be able to encode word orders that are not attested—or that Cinque predicts not to be attested—Merlo uses different merge orders, i.e. a different underlying base structures.

at the token level (where they have to predict the frequency class of individual languages), and in each case with three levels of granularity for the dependent variable: one with two values (possible, impossible), one with four (very frequent, frequent, rare, unattested), and one with seven (two levels of very frequent, two levels of frequent, two levels of rare, one for unattested). Moreover, the three analyses under consideration are compared against an uninformed baseline, whereby each word order or language is simply assigned to the most frequent class. The results of the experiments are represented in Figure 4.8.

	Naive Bayes					
	Type (24)			Token (214)		
	Two	Four	Seven	Two	Four	Seven
Cinque	88	58	42	97	87	89
Cysouw	*67*	*21*	66	*93*	90	68
Dryer	92	54	63	97	93	71
Baseline	71	50	38	97	47	28

Figure 4.8 Percentages of word orders ("Type") and languages ("Token") correctly classified by the three analyses into two, four, or seven frequency classes. Percentages in italics are below the baseline (Merlo 2015:336)

As is clear from the table in Figure 4.8, the method used by Merlo and outlined above provides a nuanced and precise picture of how well each of the three analyses—or at the very least, the vector encodings of these three analyses—succeeds in accounting for the variation in DP-internal word order across languages (and see Futrell et al. (2017) for a similar analysis, with similar results, but using Poisson regression instead of a Bayesian classifier). What is more, though, is that this methodology can also shed light on other aspects of these analyses. Recall that one of the central assumptions of a Naive Bayes classifier is that of independence between the attributes. Merlo performs the same set of experiments a second time, but this time with a classifier—an averaged weighted one-dependence estimator—that makes weaker independence assumptions between the attributes. That second classifier makes better predictions for Cinque's (and Dryer's) analysis. That means that even though every ingredient of Cinque's analysis—merge order, partial or full movement, type of pied-piping—is independently theoretically motivated, part of the variation can only be explained through an interaction of some of these properties, i.e. there is a dependence in the analysis which was not noted in Cinque's original account. In a similar vein, Merlo and Ouwayda

(2018) use linear regression to examine the 'cost' or markedness of certain syntactic operations in Cinque's analysis. In order to account for the (large) differences in frequencies between the various word orders, Cinque assumes that some movement or pied-piping operations are more marked than others and as a result there will be fewer word orders that make use of these operations in the world's languages. By using these operations as independent variables in a regression analysis, Merlo & Ouwayda can make their cost very precise as well as deduce a ranking, which can then be compared against Cinque's. The two rankings are given in (13), where the '<'-symbol means "is less costly than" and '=' means "is equally costly as".

(13) a. **Cinque**
whose picture pied-piping = partial movement < NP-movement without pied-piping < *picture of who* pied-piping < NP-subextraction = movement without NP

b. **Merlo & Ouwayda**
whose picture pied-piping < partial movement < NP-subextraction < *picture of who* pied-piping < NP-movement without pied-piping < movement without NP

The ranking that falls out from the linear regression analysis is very similar to that of Cinque. This is also confirmed by a statistical test: Kendall's Tau_b is a measure of rank correlation that allows ties, and the comparison between (13a) and (13b) yields a Kendall's Tau_b of 0.6 ($p < 0.5$). At the same time, there is a noticeable difference concerning the placement of NP-subextraction, an operation whereby NP-movement first pied-pipes other material, but then strands that material on its way to a higher DP-internal landing site. Cinque (2005:323) considers this to be an extremely marked process, while Merlo and Ouwayda's analysis shows it to be the third least costly operation in Cinque's arsenal. This shows how a quantitative approach can provide a very detailed and precise feedback loop for theoretical analyses. For instance, Cinque's reason for thinking that NP-subextraction is highly marked is that it violates the so-called Freezing Principle, which bans movement out of previously moved phrases, but Merlo and Ouwayda's results might lead one to rethink the strength of that ban (see also Abels (2009)). Merlo and Ouwayda also focus on another ingredient of Cinque's analysis, namely the base-generated order, but here their conclusions are more directly in line with Cinque's. Specifically, they examine how the proposed base order Dem-Num-A-N fares against alternatives whereby numerals are—either categorically or in a subset of the cases—merged below adjectives. The experiment reveals that an analysis based on a universal base order Dem-Num-A-N yields the best empirical fit.

Continuing on the topic of DP-internal word order, Gulordava and Merlo (2020) focus on Greenberg (1963)'s Universal 18, reproduced in (14) (Greenberg 1963:67–68).

(14) **Greenberg's Universal 18**
When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise.

Gulordava and Merlo (2020) zoom in on the position of adjectives and numerals inside the DP, and they distinguish between token-level and type-level accounts of Universal 18. The former assume there to be a dispreference for—or even a ban on—specific structural instantiations of the noun phrase, i.e. ones whereby the adjective precedes the noun while the numeral follows: [[A N] Num]. A typical exemplar of this type of analysis are FOFC-based accounts, see, e.g., Biberauer et al. (2014). Type-level analyses operate at the level of the entire language and assume there to be a bias against the co-occurrence of two language-wide properties, namely the probability of the adjective preceding the noun being higher than chance and the probability of the numeral following the noun also being higher than chance (see, e.g., Culbertson et al. (2012)). As Gulordava and Merlo (2020) point out, these two types of analyses make different predictions when pitted against corpus data. Token-based analyses predict there to be an interaction between the frequency of A<N-orders and that of N<Num-orders: the combination of the two—i.e. A<N<Num-orders—should be less frequent than expected based on the frequency of the individual orders. Put differently, it should in principle be possible for both A<N-orders and N<Num-orders to be extremely frequent in a language, as long as they do not co-occur in one and the same noun phrase. Type-level accounts, on the other hand, do not make such a prediction: whatever the probability of A<N- and N<Num-orders, that should be independent of whether adjectives and numerals occur in the same noun phrase or not. In other words, the probability of the order A<N<Num should simply be the product of the probabilities of A<N and N<Num. Gulordava and Merlo focus on Latin and Ancient Greek, two languages known for their high degree

of DP-internal word order freedom, and they show that the data favor token-based explanations of Universal 18 over type-based ones: the observed frequency of A<N<Num-orders is smaller than would be expected based on the individual frequencies of A<N- and N<Num-orders.

Van Craenenbroeck et al. (2019) and Van Craenenbroeck & Van Koppen (2023) use *k*-nearest neighbours (henceforth *k*NN) classification to test linguistic analyses of syntactic variation. Like Naive Bayes, this is a supervised learning method that takes as input a set of labeled training data and that uses this information to classify new data points. Van Craenenbroeck et al. (2019) and Van Craenenbroeck & Van Koppen (2023) use a specific implementation of cross-validation called 'leave one out', whereby each individual data point in turn serves as the new, unseen data point, while all the remaining ones constitute the training data. This means that in a data set of, say, 100 data points, the experiment is run 100 times. In determining the classification of a new data point, the *k*NN-algorithm only takes into account the attributes and classification of the *k*—a natural number—data points that are most like the new point, i.e. its nearest neighbors. So if $k = 1$, the algorithm copies the classification of the known data point that is most similar to the new one. This is particularly interesting for examining the effect of geographical proximity on language variation. Consider for example the distribution in Figure 4.9 of the phenomenon shown in (15), where a determiner and a demonstrative pronoun co-occur in the context of NP-ellipsis.

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

The map in Figure 4.9 indicates for 260 dialect locations of Dutch whether or not they feature determiner-demonstrative doubling (black dots = 'yes', transparent dots = 'no'). Suppose now that we add a point and use the *k*NN-method to classify it as a black or a transparant dot based on its geographical location. If $k = 1$, the algorithm looks at the dialect location that is geographically closest to the new point, and gives the new point the same classification as its neighbor. If $k = 3$, the three closest dialect locations are taken into account, and the new point gets the same classification as the majority of those three points.[7] In other words, the value of *k* is a measure for how widely the geographical net is

[7] In the case of a tie—which can arise when *k* is even—various implementations

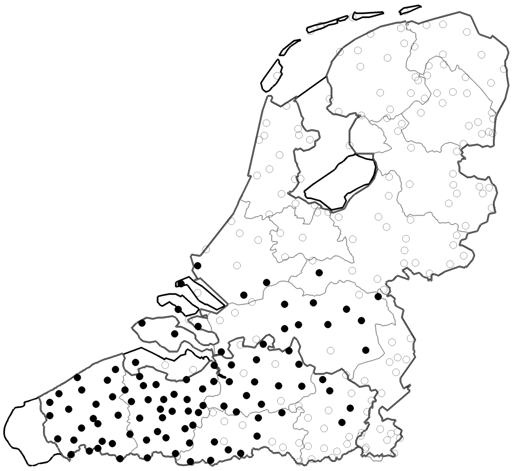


Figure 4.9 Geographical distribution of determiner-demonstrative doubling in Dutch dialects (data from Barbiers *et al.* (2006)).

cast in the search for nearest neighbors. What is interesting about kNN-classification, though, is that it can use distance measures other than Euclidean distance as well. The analyses Van Craenenbroeck et al. (2019) and Van Craenenbroeck & Van Koppen (2023) work with are parametric accounts of sets of microvariation data: they identify various parameters the setting of which determines the occurrence or non-occurrence of particular sets of linguistic phenomena. Each dialect location is thus given its own parametric analysis. Consider in this respect the hypothetical situation in the table in (16).

(16)

	param. 1	param. 2	param. 3	phenomenon X
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

This table represents a situation whereby a number of dialect locations are characterized not based on their geographical location but on the ba-

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.

sis of their parameter setting, and the relevant classification is whether or not a particular linguistic phenomenon (in this hypothetical case marked as ‘phenomenon X’) occurs in those locations or not. Once again, kNN-based classification can serve to analyze these data. Suppose the occurrence of X in dialect A is unknown. The kNN-method tries to determine that classification based on the parameter setting of dialect A. If $k = 1$ it only looks at the closest possible parameter setting, which in this case would be that of dialect B, which has the same parameter setting as A, but if $k = 2$ dialects C and D are also taken into account because their parameter setting differs from that of A in exactly one value.

Van Craenenbroeck et al. (2019) and Van Craenenbroeck & Van Koppen (2023) each use kNN-classification to test a parametric analysis of a set of Dutch dialect data and to compare it against a geography-based account. This allows them not only to determine if the analysis can pick up a signal in the data that is not purely geography-driven—thus reducing the chance of correlations in the data being due to language contact, shared history, etc.; see subsection 4.2.1 above for discussion—but also to see if there is potential complementarity between the two approaches. Consider for example the maps in Figure 4.10.

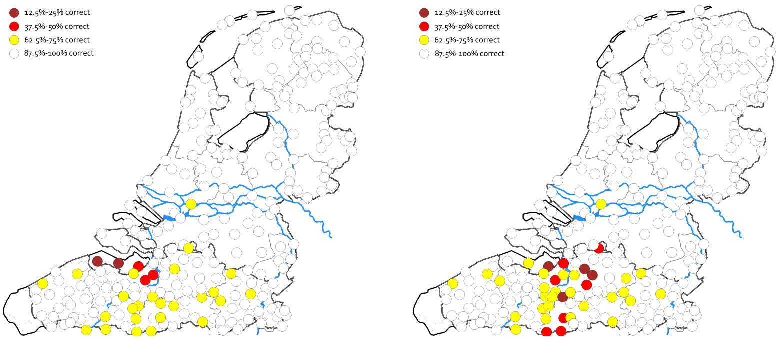
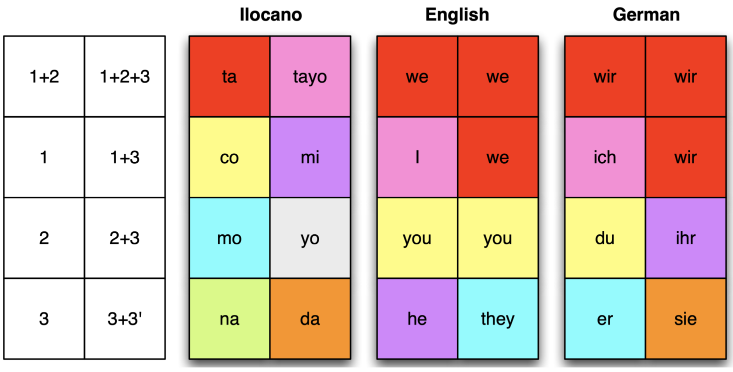


Figure 4.10 Percentage of correct predictions per dialect location for the location-based account (left) and the parameters-based one (right) (Van Craenenbroeck & Van Koppen (2023:11))

These maps visualize the percentage of phenomena correctly classified per dialect location based on geographical location (left) and the parametric account (right) respectively. As is clear from the positioning of the yellow and especially the red dots, the two accounts differ in where they make wrong predictions. While the parameter-based analysis struggles with the transition between dialect areas—transitions which are known to be gradual, but which a parameter-based account by definition classifies as sharp and categorical—the geography-based analysis performs poorly near the country border, where geographical proximity is not always a good indicator for linguistic similarity.

So far we have looked at accounts whereby a qualitative theoretical analysis is formalized in such a way that it can be used as input for a quantitative analysis, which makes it possible to test the empirical fit of those theoretical accounts in a precise and detailed way, and can even lead us to reassess certain operations and assumptions that are part of the account. Now we discuss a number of accounts that are quantitative first, but that have clear implications and consequences for possible theoretical accounts. A first example is Sauerland and Bobaljik's (2013) discussion of the typological distribution of syncretism across person paradigms. Consider in this respect the paradigms in Figure 4.11.



		Ilocano		English		German	
1+2	1+2+3	ta	tayo	we	we	wir	wir
1	1+3	co	mi	I	we	ich	wir
2	2+3	mo	yo	you	you	du	ihr
3	3+3'	na	da	he	they	er	sie

Figure 4.11 Person paradigms for Ilocano, English, and German (from Sauerland and Bobaljik (2013:37))

This figure contains the full eight-cell paradigms of personal pronouns for Illocano, English, and German. It distinguishes not just first, second, and third person, but also first person inclusive (1+2), and it makes a systematic singular–plural (or minimal–non-minimal) distinction for all persons (represented here as '+3', i.e. the addition of one or more non-participants). As is clear from the colored tables, Ilocano uses a different pronominal form for each of the eight cells, while German distinguishes six different forms, and English only five. In other words, English and German display a certain degree of syncretism in their pronominal paradigm, while Ilocano does not. Syncretism has become an intensely researched and hotly debated topic in recent years (see e.g. Baerman et al. (2005)) and one of the central questions in the literature is to what extent syncretism is systematic or accidental. If systematic, patterns of syncretism can be informative about the underlying feature hierarchy (Caha 2009, Bobaljik 2012) or the workings of the syntax-phonology interface (e.g. the presence of Impoverishment rules in Distributed Morphology), while accidental homophony should be treated in much the same way as cases of lexical homophony—such as the fact, say, that the Dutch word *vorst* can signify both a monarch and frost—i.e. as the result of (co)incidental (and often sound-related) historical developments.

The problem, however, is distinguishing between the two: while some cases—like the Dutch *vorst*-example—are clearly accidental, the majority of the cases is less clear-cut and their characterization as systematic or accidental often depends on one's theoretical assumptions. The main innovation of Sauerland and Bobaljik (2013) is that they propose a mathematically precise way of defining accidental homophony and thus of making a distinction between systematic and accidental homophony. They propose that accidental homophony should be assumed to be a random event in the statistical sense of the term, i.e. as a random factor with a constant probability across all cases. This means that we can start comparing the empirical fit of morphosyntactic analyses of syncretism patterns based on the degree to which the instances they identify as being accidental homophony bear the expected statistical signature. Sauerland & Bolbaljik call this method Syncretism Distribution Modeling, and while their paper is mostly a proof of concept intended to show how the method works, they do use it to rule out two extreme null hypotheses: one that assumes all homophony is accidental, and one that assumes all of it to be systematic. It is clear how in future work Syncretism Distribution Modeling could be used to test and (dis)confirm actual analyses of syncretic paradigms across languages.

The final case study we want to discuss here is that of Mahowald et al. (2021). The phenomenon they focus on is concord in the nominal domain. Based on a data set of nominal concord in 174 languages from 105 families (Norris 2019, 2020), they first train a hierarchical Bayesian model which models the occurrence of concord based on the type (gender, number, case, definiteness) and locus (noun, adjective, demonstrative) of the concord, a number of additional properties of the languages (e.g. the presence of case marking or DP-internal word order), and random effects for language, family, and area. The fit of the model is assessed using five-fold cross-validation—see above for discussion—and it yields an accuracy of 87% to 92% (depending on how the test data set is created). In a next step, Mahowald et al. use posterior draws from this model to explore how concord varies within and across languages. An intuitive way to understand this would be to compare it to a model that simulates coin flips of a fair coin, except instead of heads or tails, the model is predicting the presence or absence of concord, and it does so based on a large set of parameters rather than the one in two chance of landing on either side in the case of a fair coin. Mahowald et al. repeatedly simulate data for (thousands of) hypothetical new languages from hypothetical language families and language areas, and then examine the properties of this new data set. For example, is there an implicational relation between number and gender concord? How many of the languages generated have a particular type of concord? etc. The main advantage of this approach is that it allows one to make generalizations about concord that go beyond areal and family-specific effects, and while Mahowald et al. do not discuss specific theoretical analyses of concord that are or are not compatible with their findings, it is not hard to see how several of their results could have a direct bearing on theoretical accounts. To name but one, they show that there is no strong correlation between DP-internal word order and the presence of concord—unlike what is suggested by Greenberg's (1963:95) Universal 40—a finding which would be directly relevant for accounts that tie word order changing movement operations to the presence of overt morphology.

This concludes our overview of case studies that integrate both quantitative and qualitative components into their analysis of syntactic variation. In the next section we move away from these concrete examples and provide some more general considerations.

4.3 General considerations

This section addresses two related topics. First, we discuss some of the advantages of adopting a combined quantitative-qualitative approach towards studying syntactic variation. Then, we show how looking at variation data from such an integrated perspective raises fundamental theoretical issues about the nature of variation itself and the properties of the grammar responsible for generating said variation.

For the formal-theoretical linguist interested in studying syntactic variation, there are clear advantages to adding a quantitative perspective to their research. Generally speaking, it adds more solid ground and a higher degree of precision to such analyses, not only because a quantitative analysis typically entails looking at a large data set, but also because the integration of (components of) abstract theoretical analyses into a quantitative model requires a formulation of those theories that is unambiguous, formally precise, and falsifiable (see for example the encoding of Cinque's movement-based analysis of DP-internal word order in (12) and see also Guardiano et al. (2016:96) on this point). Moreover, as was discussed in subsection 4.2.3, quantitative analyses can provide objective measures against which to gauge the (empirical) success of a formal analysis, to compare analyses against one another, or to compare analyses against a previously established baseline. In each case, the explanatory force of the analysis is strengthened and its plausibility increased. At a more fundamental level, though, adopting a quantitative perspective can lead theoretical linguists to take into consideration data types and patterns that they would otherwise have overlooked. A good example of this is frequency data. Frequency is often considered—by generative linguists in particular—to be exclusively part of performance/E-grammar, not of competence/I-grammar, and as a result not worthy of serious linguistic investigation. Many of the accounts highlighted in the previous section, however, argue that inherently gradient, and hence quantitative frequency data can contain a grammatical signal as well. Consider for example Samo and Merlo's (2021) discussion of intervention effects in clefts. While both subject and object clefts are judged to be grammatical by native speakers of English, French, and Italian, there are clear differences in corpus frequency between the two, and Samo & Merlo show convincingly that the reduced frequency of object clefts is due to a Rizzian intervention effect caused by the object having to move over the subject.[8] This is an intervention effect that does not rise to the level of full ungrammaticality, but that is present in the data nonetheless, and it is crucially the quantitative analysis that brings out these kinds of patterns and generalizations.

The advantages of adopting an integrated quantitative-qualitative approach extend in the other direction as well: a quantitative or computational linguist can benefit greatly from incorporating insights from formal theorizing into their research. A solid grasp of formal theoretical principles and results can help better characterize both the input and

[8] We're simplifying the discussion somewhat here, as Samo and Merlo (2021) not only take into account the mere fact of intervention, but also the nature of the intervener, with featurally more similar interveners (correctly) predicted to show stronger intervention effects. See the original paper for details.

the output of a quantitative analysis. Recall for example Burnett et al.'s (2018) discussion of negative quantifiers and negative polarity items from subsection 4.2.2. They show how encoding the raw data in terms of which DPs are inside the VP and which ones have undergone object shift—a highly technical and inherently theoretical classification—yields better results than looking at surface phenomena such as the type of main verb in the clause. In other words, theoretical insights yield better predictor variables, which in turn lead to more successful statistical models. Relatedly, formal linguists can also help interpret the outcome of a statistical analysis, especially when it yields a high volume of results. As an example, consider Spruit's (2008) analysis of the data from the Syntactic Atlas of the Dutch Dialects (SAND). One of the techniques he uses is association rule mining, whereby an algorithm searches for implicational relationships—*if. . . then*-rules—between (groups of) variables. When applied to the 485 syntactic variables in the SAND-database, this technique yields no less than 56,267,729 rules that have an accuracy of at least 90 percent. This suggests that in addition to purely quantitative measures such as coverage and accuracy, we also need more qualitative measures to be able to assess which association rules are interesting or relevant, and this is where theoretical linguistic insights might be of use. In the words of Spruit (2008:106) himself, the quantitative method "will require extensive consultation with syntactic theorists to meaningfully interpret the data."

In short, it is clear that a collaboration between qualitatively oriented theoretical linguists and quantitative-statistical linguists can be hugely mutually beneficial. At the same time and at a more general level, the kind of integrated research just sketched by its very nature raises issues and questions that are fundamental to our understanding of syntactic variation. For example, the types of analyses discussed in the previous section raise—and in many cases provide partial answers to—questions such as the following:

(17) a. What is the relationship between language-internal variation and typological variation?
b. What is the relationship between syntactic variation and phonological variation?
c. What is the relationship between native speaker judgments and corpus frequencies?

The first question was already touched upon earlier in this section, when we discussed Samo and Merlo (2021). Their study is but one of a whole series pointing out that contrasts that lead to ungrammaticality across languages often surface as statistical preferences within one language, or in the words of Bresnan et al. (2001): "soft constraints mirror hard constraints". Given that it seems to be the exact same grammatical principles at play in both cases—again, see Samo and Merlo (2021) for a particularly clear illustration of this, but also, e.g., Samardžić and Merlo (2018)—this shows that gradience and variability have to be built into the grammatical pipeline at some point, be it in the grammar itself (Yang 2002, Manning 2003, Bresnan et al. 2007), or at a transition point to one of the interfaces (see for example Adger (2006)).

The second question in (17) is occasionally discussed in the quantitative literature, and although there is no definitive answer to it, the growing consensus seems to be that the two types of variation pattern differently. For example, Spruit (2008:86) finds only a modest correlation ($r = 0.35$) between the data from the syntactic and phonological atlases of the Dutch dialects, and Birkenes and Fleischer (2022) in their quantitative study of Hessian dialects conclude that the two types of variation have a different geographical signature: "syntax is more prone to nonareal variation. Similar syntactic distributions are, to some extent, areally discontinuous. In contrast, the choice between phonological variants seems to be more of a categorial nature." (Birkenes and Fleischer 2022:157) (and see also Scherrer and Stoeckle (2016) for a similar conclusion). If such results are corroborated in future research, they might prove informative about the nature of syntactic variation. Recalling the discussion in subsection 4.2.1, the lack of a clear geographical signal in syntactic variation—especially when such a signal *is* present in phonological variation between the same varieties—might be an indication that syntactic variation is grammar-driven and not solely due to extra-grammatical factors such as language contact, a shared history, etc.

The third question in (17) is more methodological in nature, although it also touches on fundamental theoretical issues like the competence-performance distinction alluded to above. As far as we can tell, it is also one where the jury is still out, with some studies (see Bresnan (2007) for a clear example) finding that native speaker intuitions closely mirror corpus probabilities, and others finding floor or ceiling effects in corpus frequencies in contexts where native speaker judgments provide a more nuanced and varied picture (see e.g. Bader and Häussler (2010) and Cavirani-Pots (2020)).

In summary, adopting an integrated quantitative-qualitative approach towards studying syntactic variation is not only mutually beneficial to both parties involved, it also raises fundamental theoretical issues that go to the heart of variation itself.

4.4 Conclusion

In this chapter we have focused on quantitative approaches to studying syntactic variation. We have deliberately narrowed down that topic to approaches that pursue a combined quantitative-qualitative approach, integrating components both from the formal-theoretical and the computational-statistical tradition, because we feel it is an area of great potential and promising prospects. An integrated approach of this type is mutually beneficial to linguists of both persuasions, and it has the potential of substantially deepening our understanding of syntactic variation. By its very nature, this is an endeavor that benefits from intense collaboration, as it requires its practitioners to be well-versed in both sides of the equation: in-depth knowledge and understanding of the theoretical analyses is necessary in order to implement and model them, and at the same time deep insight into and understanding of the quantitative techniques is necessary to set up the experiments, choose and apply the methodology, interpret the results, control for possible interfering factors, etc. It is this spirit of collaboration we hope will become the dominant paradigm of the future in syntactic variation research.

References

Abels, Klaus. 2009. Some implications of improper movement for cartography. Pages 325–360 of: van Craenenbroeck, Jeroen (ed), *Alternatives to cartography.* Berlin: Mouton de Gruyter.

Adger, David. 2006. Combinatorial variability. *Journal of Linguistics*, **42**(3), 503–530.

Bader, Markus, and Häussler, Jana. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics*, **46**(2), 273–330.

Baerman, Matthew, Brown, Dunstan, and Corbett, Greville. 2005. *The syntax-morphology interface: A study of syncretism.* Cambridge: Cambridge University Press.

Barbiers, Sjef. 2005. Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics. Pages 233–264 of: Cornips, Leonie, and Corrigan, Karen P. (eds), *Syntax and variation. Reconciling the biological and the social.* Current issues in linguistic theory, vol. 265. Amsterdam: John Benjamins.

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

Barbiers, Sjef, et al. 2006. *Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND).* Meertens Institute. www.meertens.knaw.nl/sand/.

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

Biberauer, Theresa, Holmberg, Anders, and Roberts, Ian. 2014. A syntactic universal and its consequences. *Linguistic Inquiry*, **45**(2), 169–225.

Birkenes, Magnus Breder, and Fleischer, Jürg. 2022. Syntactic vs. phonological areas: a quantitative perspective on Hessian dialects. *Journal of Linguistic Geography*, **9**(2), 142–161.

Bobaljik, Jonathan. 2012. *Universals in Comparative Morphology: Suppletion, Superlatives, and the Structure of Words.* Cambridge, Mass.: MIT Press.

Bosch, Anna R. K. 2006. Scottish Gaelic dialectology: A preliminary assessment of the Survey of the Gaelic Dialects of Scotland. *Lingua*, **116**(11), 2012–2022.

Bresnan, Joan. 2007. Is syntactic knowledge pobabilistic? Experiments with the English dative alternation. Pages 75–96 of: Featherston, Sam, and Sternefeld, Wolfgang (eds), *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton de Gruyter.

Bresnan, Joan, Dingare, Shipra, and Manning, Chris. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. Pages 13–32 of: Butt, Miriam, and Holloway King, Tracy (eds), *Proceedings of the LFG01 Conference*. Stanford: CSLI Publications.

Bresnan, Joan, Deo, Ashwini, and Sharma, Devyani. 2007. Typology in variation: a probabilistic approach *to be* and *n't* in the Survey of English Dialects. *English Language and Linguistics*, **11**(2), 301–346.

Burnett, Heather, Koopman, Hilda, and Tagliamonte, Sali. 2018. Structural explanations in syntactic variation: The evolution of English negative and polarity indefinites. *Language Variation and Change*, **30**, 83–107.

Caha, Pavel. 2009. *The Nanosyntax of Case*. Ph.D. thesis, University of Tromsø, Tromsø.

Cavirani-Pots, Cora. 2020. *Roots in progress. Semi-lexicality in the Dutch and Afrikaans Verbal Domain.* Ph.D. thesis, KU Leuven.

Cinque, Guglielmo. 2005. Deriving Greenberg's Universal 20 and Its Exceptions. *Linguistic Inquiry*, **36**(3), 315–332.

Cornips, Leonie, and Corrigan, Karen. 2005. Toward an integrated approach to syntactic variation: A retrospective and prospective synopsis. Pages 1–30 of: Cornips, Leonie, and Corrigan, Karen (eds), *Syntax and variation. Reconciling the biological and the social.* Amsterdam/Philadelphia: John Benjamins.

Culbertson, Jennifer, Smolensky, Paul, and Legendre, Géraldine. 2012. Learning Biases Predict a Word Order Universal. *Cognition*, **122**(3), 306–329.

Cysouw, Michael. 2010. Dealing with diversity: towards an explanation of NP word order frequencies. *Linguistic Typology*, **14**(2), 253–287.

De Troij, Robbert, Grondelaers, Stefan, and Speelman, Dirk. 2023. Natiolectal Variation in Dutch Morphosyntax: A Large-Scale, Data-Driven Perspective. *Journal of Germanic Linguistics*, **35**(1), 1–68.

de Villiers, Jill, and Roeper, Tom (eds). 2011. *Handbook of generative approaches to language acquisition.* New York: Springer.

Dryer, Matthew. 2006. *The order demonstrative, numeral, adjective and noun: an alternative to Cinque.* Ms. University of Buffalo.

Dryer, Matthew. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language*, **94**(4), 798–833.

Evers, Arnold. 1975. *The Transformational Cycle in Dutch and German.*

Futrell, Richard, Levy, Roger, and Dryer, Matthew. 2017. A Statistical Comparison of Some Theories of NP Word Order. *CoRR*, **abs/1709.02783**.

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, Joseph (ed), *Universals of language.* Cambridge, Massachusetts: MIT Press.

Guardiano, Cristina, Michelioudakis, Dimitris, Ceolin, Andrea, Longobardi, Giuseppe, Irimia, Monica, Radkevich, Nina, Sitaridou, Ioanna, and Silvestri, Giuseppina. 2016. South by Southeast. A syntactic approach to Greek and Romance microvariation. *L' Italia Dialettale*, **LXXVII**, 95–166.

Guardiano, Cristina, Longobardi, Giuseppe, Cordoni, Guido, and Crisma, Paola. 2020. Formal Syntax as a Phylogenetic Method. Pages 145–182 of: Janda, Richard D., Joseph, Brian D., and Vance, Barbara S. (eds), *The Handbook of Historical Linguistics.* John Wiley & Sons, Ltd.

Guardinao, Cristina, and Longobardi, Giuseppe. 2005. Parametric Comparison and Language Taxonomy. Pages 149–174 of: Batllori, Montserrat, Hernanz, Maria-Lluïsa, Picallo, Carme, and Roca, Francesc (eds), *Grammaticalization and Parametric Variation.* Oxford University Press.

Gulordava, Kristina, and Merlo, Paola. 2020. Computational Quantitative Syntax: The Case of Universal 18. Pages 109–132 of: Vogel, Irene (ed), *Romance Languages and Linguistic Theory 16 Selected papers from the 47th Linguistic Symposium on Romance Languages (LSRL), Newark, Delaware.* Amsterdam/Philadelphia: John Benjamins.

Haegeman, Liliane, and van Riemsdijk, Henk. 1986. Verb Projection Raising, Scope, and the Typology of Verb Movement Rules. *Linguistic Inquiry*, **17**(3), 417–466.

Halle, Morris, and Marantz, Alec. 1993. Distributed morphology and the pieces of inflection. Pages 111–176 of: Hale, Ken, and Keyser, Jay (eds), *The View from Building 20.* Cambridge, MA: MIT Press.

Haspelmath, Martin. 2008. Parametric versus functional explanations of syntactic universals. Pages 75–107 of: Biberauer, Theresa (ed), *The Limits of Syntactic Variation.* Amsterdam: John Benjamins.

Iosad, Pavel, and Lamb, William. 2020. Dialect variation in Scottish Gaelic nominal morphology: a quantitative study. *Glossa: a journal of general linguistics*, **5**(1), 130.

Jeszenszky, Péter, Stoeckle, Philipp, Glaser, Elvira, and Weibel, Robert. 2017. Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*, **5**(2), 1–23.

Kayne, Richard S. 1998. Overt vs. covert movement. *Syntax*, **1**(2), 128–191.

Kazakov, Dimitar, Cordoni, Guido, Ceolin, Andrea, Irimia, Monica, Kim, Shin-Sook, Michelioudakis, Dimitris, Radkevich, Nina, Guardiano, Cristina, and Longobardi, Giuseppe. 2017. Machine Learning Models of Universal Grammar Parameter Dependencies. Pages 31–37 of: Zervanou, Kalliopi, Osenova, Petya, Wandl-Vogt, Eveline, and Cristea, Dan (eds), *Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP 2017.* Varna: INCOMA Inc.

Ledgeway, Adam, and Roberts, Ian (eds). 2017. *The Cambridge Handbook of Historical Syntax.* Cambridge: Cambridge University Press.

Longobardi, Giuseppe. 2003. Methods in parametric linguistics and cognitive history. Pages 101–138 of: Pica, Pierre, and Rooryck, Johan (eds), *Linguistic Variation Yearbook.* Amsterdam: John Benjamins.

Longobardi, Giuseppe. 2018. Principles, parameters, and schemata: a radically underspecified UG. *Linguistic Analysis*, **41**(3–4), 517–558.

Longobardi, Giuseppe, and Guardiano, Cristina. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, **119**, 1679–1706.

Mahowald, Kyle, Jurafsky, Dan, and Norris, Mark. 2021. Concord begets concord: a Bayesian model of nominal concord typology. *Proceedings of the Linguistic Society of America*, **6**(1), 541–555.

Manning, C. 2003. Probabilistic syntax. Pages 298–341 of: Bod, Rens, and Jannedy, Stefanie (eds), *Probabilistic Linguistics.* Cambridge, Mass.: MIT Press.

Merlo, Paola. 2015. Predicting word order universals. *Journal of Language Modeling*, **3**(2), 317–344.

Merlo, Paola. 2016. Quantitative computational syntax: some initial results. *Italian Journal of Computational Linguistics*, **2**(1), 11–29.

Merlo, Paola, and Ouwayda, Sarah. 2018. Movement and structure effects on Universal 20 word order frequencies: A quantitative study. *Glossa: a journal of general linguistics*, **3**(1), 1–35.

Norris, Mark. 2019. *A typological perspective on nominal concord.* Archived data: https://shareok.org/handle/11244/320354.

Norris, Mark. 2020. *Typology of nominal concord.* Archived data: https://doi.org/10.17605/OSF.IO/TM49Q.

Pescarini, Diego. 2019. Microvariation and Microparameters. Some Quantitative Remarks. *Quaderni di Linguistica e Studi Orientali / Working Papers in Linguistics and Oriental Studies*, **5**, 255–277.

Pescarini, Diego. 2022. A quantitative approach to microvariation: negative marking in central Romance. *Languages*, **7**(87), 1–20.

Poletto, Cecilia. 2003. Leopard Spot Variation: What Dialects Have To Say About Variation, Change and Acquisition. *Studia Linguistica*, **67**(1), 165–183.

Rizzi, Luigi. 1986. Null objects in Italian and the theory of *pro*. *Linguistic Inquiry*, **17**, 501–558.

Samardžić, Tanja, and Merlo, Paola. 2018. Probability of external causation: an empirical account of cross-linguistic variation in lexical causatives. *Linguistics*, **56**(5), 895–938.

Samo, Giuseppe, and Merlo, Paola. 2019. Intervention effects in object relatives in English and Italian: a study in quantitative computational syntax. Pages 46–56 of: *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019).* Paris, France: Association for Computational Linguistics.

Samo, Giuseppe, and Merlo, Paola. 2021. Intervention effects in clefts: a study in quantitative computational syntax. *Glossa: a journal of general linguistics*, **6**(1), 1–39.

Sauerland, Uli, and Bobaljik, Jonathan David. 2013. Syncretism Distribution Modeling: Accidental Homophony as a Random Event. Pages 31–53 of: Goto, Nobu, Otaki, Koichi, Sato, Atsushi, and Takita, Kensuke (eds), *The Proceedings of GLOW in Asia IX*.

Scherrer, Yves, and Stoeckle, Philipp. 2016. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectología et Geolinguistica*, **24**(1), 92–125.

Spruit, Marco René. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. Ph.D. thesis, Universiteit van Amsterdam.

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt, and Kortmann, Bernd. 2009. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua*, **119**(11), 1643–1663.

Tottie, Gunnel. 1991a. Lexical diffusion in syntactic change: Frequency as a determinant of linguistic conservatism in the development of negation in English. Pages 439–468 of: Kastovsky, Dieter (ed), *Historical English syntax*. Berlin: Mouton de Gruyter.

Tottie, Gunnel. 1991b. *Negation in speech and writing*. San Diego: Academic Press.

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

van Craenenbroeck, Jeroen, and van Koppen, Marjo. 2023. Parameters and language contact: Morphosyntactic variation in Dutch dialects. *Catalan journal of linguistics*, **22**, 1–25.

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

Wood, Jim, and Zanuttini, Raffaella. 2018. Datives, data and dialect syntax in American English. *Glossa: a journal of general linguistics*, **3**(1), 1–22.

Yang, Charles D. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.