The Routledge Companion to Sociolinguistics

Edited by
Carmen Llamas, Louise Mullany
and Peter Stockwell

2007

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

4

TECHNIQUES OF ANALYSIS

II MORPHOSYNTACTIC VARIATION

JENNIFER SMITH

Since Labov's first groundbreaking study of Martha's Vineyard in the early 1960s, **phonetic** variation has been predominant in the field of **variationist sociolinguistics**. However, **morphosyntactic** variation has also garnered considerable attention, with a plethora of studies into variables which are common to a large number of **dialects**. These include *was/were* alternation as in (1) below; **negative concord**, as in (2); verbal *–s* as in (3); **non-standard** verb forms as in (4); **copula** deletion as in (5) and **quotative** *be like*, as in (6):

(1) The coppers let them go to see if they *was* the bastards. (Cheshire 1982: 44)
(2) I *ain't* got *no* money. (Howe and Walker 2000: 111)
(3) Her gives me a hug and a kiss when I *comes* in and one when I go. (Godfrey and Tagliamonte 1999: 89)
(4) My two brothers, they never *fighted*, you know. (Eisikovits 1987: 127)
(5) I feel like I ø fourteen. (Weldon 2003: 7)
(6) I'*m like* 'Joe, how's the truck? And he'*s like* 'Oh, Clarky, man, I fucked my truck up!' (Tagliamonte and Hudson 1999: 148)

Morphosyntactic variation is not confined to competition between dialect and **standard** forms. Variation occurs in all spoken varieties, even in those which can be considered to be fairly standard. For example, negative versus auxiliary contraction, as in (7); deontic **modality**, as in (8); *that* complementizer, as in (9); use of intensifiers, as in (10); relative clause markers, as in (11).

(7) He'*ll* not be better again Margaret, no . . . And you *won't* have the same interest. (Tagliamonte and Smith 2002)
(8) If she goes out she *must* have her chair, *got to* take her chair and a oxygen cylinder this height. (Tagliamonte and Smith in press)
(9) I think *that some of his family* would be the same. I *think* Ø she was lucky to get him. (Tagliamonte *et al.* to appear)
(10) It was a *really* old building . . . it was a *very* old rambling mess of a building. (Ito and Tagliamonte 2003: 269)
(11) The last meeting Ø we had in that church.
And they used the old nets which we *would* call strabbles.
Then there were a word *that* I couldn't get summat to rhyme with.
(Tagliamonte *et al.* in press)

These examples demonstrate that morphosyntactic variation is not some peripheral phenomenon confined to a handful of obscure varieties but instead is pervasive in everyday speech in every **variety**. Despite this, Labov (1991: 277) comments that '[s]yntactic change is an elusive process as compared to sound change'. This chapter aims to provide some insights into this process, setting out some of the steps involved in large-scale **quantitative** analyses of variable grammatical forms. I concentrate on issues of **transcription**, choice of **linguistic variable** for analysis, identifying the envelope of variation, coding, and statistical analysis.

TRANSCRIPTION

Although not a necessary step in the analysis, the task of identifying and analysing variable morphosyntactic forms is greatly aided by full *transcription* of the interview data. This is an extremely time-consuming initial stage, but one which is worth the effort, as the data can be mined for a number of variables without the need to return to the original recordings every time another variable is analysed.

The trick in the transcription of data for morphosyntactic analysis is to achieve a fine balance between level of detail and accessibility. A full phonetic transcription is unnecessary and, in general, phonetic or **phonological** processes are represented in standard orthography. Hence there should be no attempt to represent every nuance of speech with the inclusion of pseudo-phonetic representations and idiosyncratic spellings which makes the transcription incomprehensible (see also Macaulay 1991: 282) as in text A.

Text A

Ah hink thit it's e best hing ahv ivir saa. Thir wiz hunners o fouk ere, even tho' the tickets cos' ten powin. Ah'v got a lo' o' rispek fur a at people 'it made it happin.

On the other hand, all pertinent grammatical variation should be preserved, whether it conforms to 'standard' rules or not. Although this may seem like an extremely obvious statement to make, text B exhibits a common problem in the transcription phase: the speech has been largely standardized to what the hearer *thinks* s/he hears.

Text B

I think that it's the best show I've ever seen. There were hundreds of folk there, even though the tickets cost ten pounds. I've got a lot of respect for all those people who made it happen.

Text C is the actual words spoken during the interview, where the differences are italicized. Notice that many of the actual 'mistakes' in text B are related to variables which are so common that they are often below the level of conscious awareness – 'I think that/ø,' 'There *were/was* hundreds of folk.'

Text C

I think ø it's the best show I've ever *saw*. There *was* hundreds of folk there, even though the tickets cost ten *pound*. I *ø got* a lot of respect for all *them* people *that* made it happen.

Just how much detail goes into the transcription is in the end up to the researcher, as the final product should be guided by the goals of the study.

CHOICE OF LINGUISTIC VARIABLE

As demonstrated in the introduction, spoken data reveal a wealth of variation at the grammatical level. However, not all are ideal candidates for study. The most important issue in the choice of variable is *frequency*: morphosyntactic variables tend on the whole to be much less recurrent than phonetic variables, which can be a problem for quantitative analysis. As a guide to minimum contexts of use for quantitative analysis, Labov (1966: 181) suggests ten to twenty instances per speaker in the data, while Guy (1980) suggests more than thirty (but see further Britain 1999).

A second issue which must be dealt with in the initial stages of variable selection is functional equivalence (see Lavandera 1978; Romaine 1982; Cheshire 1987), where the differing **variants** should be 'alternat[iv]e ways of saying "the same" thing' (Labov 1972b: 118). This criterion can be easily satisfied with most phonetic variables (see Chapter 3). With many morphosyntactic variables, the same can apply, e.g. existential agreement as in (12):

(12) There *are* elephants at the party . . . there*'s* jelly sweeties for you. (Smith 2003–05)

However, there are other cases of so-called 'higher-level' variables where **semantic/pragmatic** as well as **syntactic** differences may also need to be taken into account. This is demonstrated by the 'hot news' *after perfect* in Irish English, as in (13), where the construction signals a very recent action, a meaning which the **Standard English** example in (14) does not capture:

(13) One of the farls was after breaking (Corrigan 1997: 160).
(14) One of the farls had broken.

Thus in many cases it is crucial to take into account not only the surface forms but their pragmatic inferences as well in deciding what is really equivalent (see also Milroy and Gordon 2003). This then allows the researcher to set out the envelope of variation, as discussed in the next section.

CIRCUMSCRIPTION AND EXTRACTION

Circumscription of the variable context or the envelope of variation is a major part of the analysis. In other words, what should be included in the count and what should not. These decisions come from two sources – the literature available on the forms under study and the researcher's own observations of the data. Wolfram (1979: 46) provides a perfect example of delimiting the envelope of variation in his study of *a*-prefixing as in (15):

(15) He came *a*-running down there.

He began with Krapp's (1925: 268) observation that *a*-prefixing could occur with 'every present participle'. However, closer examination of his particular data set revealed that the contexts in which this variant could occur was far more *circumscribed*. For example, the affix could not appear on an adjective such as **a-shocking*, nor verbs which did not begin with a stressed syllable (**a-repeating*).

Another example is negative concord to indeterminates following the verb, including plural NPs, as in (16), indeterminates such as *nothing* and *no one*, as in (17), and indefinite singulars, as in (18):

(16) There *wasn't no* lights. (Cheshire 1982: 65)
(17) We *never* had *nothing*. (Feagin 1979: 229)
(18) She *wasn't no* old cripple woman. (Howe and Walker 1995: 63)

Crucially, not all dialects show the full range of variability. For example, in my own analysis (Smith 2001), indefinite singulars are not a context for negative concord, thus including these in the count would have skewed the results. This is not to suggest that this is a trivial point to be ignored: why some dialects allow variation in some contexts while others do not is a crucial finding which plays a fundamental role in the final interpretation of findings.

In the examination of quotative markers in British and Canadian English, Tagliamonte and Hudson (1999) include *think* as a quotative, although in their data it is only ever used with internal dialogue, as in (19).

(19) And I *was thinking*, 'Well, surely they can all get on.' I *thought*, 'Right, OK.' (Tagliamonte and Hudson 1999: 148)

Said, on the other hand, as in (20), and many other quotative verbs are used only with direct speech:

(20) And she *said*, 'Would you like me to phone?' And I *said*, 'Don't do that 'cos Dad'll be furious!' (Tagliamonte and Hudson 1999: 148)

However, *be like* is used for both direct speech, as in (21), and internal dialogue, as in (22). Therefore, in order to account for the entire quotative system, they had to include all quotatives.

(21) She's *like* 'Right, you know we're taking you out.' (Tagliamonte and Hudson 1999: 147)
(22) And I'm like 'Oh my God, oh my God, oh my God.' I was having a heart attack. (Tagliamonte and Hudson 1999: 157)

What to include and what not to include can seem like a minefield. However, it is often the case that the researcher does not have all the answers in advance. While much of the groundwork can be carried out before extraction of the variant forms through literature sources and observation of the data, exclusion and inclusion are an on-going process. As Labov (1969: 728) observes, 'even the simplest type of counting raises a number of subtle and difficult problems. The final decision as to what to count is actually the final solution to the problem at hand. This decision is approached only through a long series of exploratory manoeuvres.'

The next step in the analysis is extraction of *all* contexts where a variant could potentially appear, in line with the 'Principle of Accountability' (Labov 1972b: 72). In other words, where a particular variant does not appear is just as important as where it does. Therefore in the case of non-standard *was* in the Buckie dialect from north-east Scotland (Smith 2000), all standard *were* contexts are included, whether they appear with *was* or with *were*, as in (23–7):

(23) They *were* all in Gaelic.
(24) *Was* you home?
(25) The plans *was* drawn up,
(26) We *wasna* actually gan thegither.
(27) There *were* four of us gied away with her to the blueberries.

The data may be extracted automatically using a concordance (e.g. Rand and Sankoff 1990) or done manually. In many cases, extraction relies on both automatic and manual extraction. For example, in the case of quotatives, it is simple to search for lexical items such as *said* and *thought*, but what about *be like*? *Like* is multifunctional: it can be a verb, a suffix, a **discourse marker** and a conjunction. Thus this is a case where the researcher must decide on which *likes* are quotatives and which are not.

Once all possible occurrences of use have been extracted, the data are ready to be *coded*.

CODING AND STATISTICAL ANALYSIS

The first stage in the statistical analysis is to count the number of tokens overall, and the proportion of different variants within these instances of use. Numbers of tokens can literally be thousands: fortunately there are computer programs which can calculate the numbers. A range of programs exist for the analysis of variation in speech: Goldvarb (Rand and Sankoff 1990) and Varbrul (Pintzuk 1988) have been used extensively in sociolinguistic research, as they are designed to deal with the types of often 'messy' data from naturally occurring talk (as opposed to

experimental data, which are highly controlled). The formats in which programs can 'read' the data differ but as it cannot 'read' straight sentences the researcher has to 'tell' the computer various pieces of information by *coding* the data. Here Goldvarb (Rand and Sankoff 1990) is used for exemplification purposes. The first piece of vital information is in the bracketed column to the left: what variant is used in the actual utterance. In the case of *was/were*, R is used to signal *were* and S *was* (the choice of code is arbitrary). This information tells the computer that in (23′) the form is *were*, but in (24′) it is *was*.

(23′) (R) They *were* all in Gaelic.
(24′) (S) *Was* you home?
(25′) (S) The plans *was* drawn up.
(26′) (S) We *wasna* actually gan thegither.
(27′) (R) There *were* four of us gied away with her to the blueberries.

Distributional analysis

Once all the occurrences of the variable have been coded for whether they appear with *was* and *were*, we are in a position to establish (1) how many occurrences of the use of the variable under study are in the data and (2) the different numbers of variants that make up these occurrences. These initial figures are known as the *overall distributions* and are normally the first set of results reported.

Table 4.1 shows the overall distribution of *was* and *were* in standard *were* contexts (see Smith 2000). These figures establish that the variable is frequent in the data and shows robust variability between the two forms, that is, both variants are present in substantial numbers in the data, making it a good candidate for quantitative analysis.

Table 4.1 Overall distribution of *was* in *were*

Word	*No.*	*%*
was	628	46
were	723	54
Total	1,351	100

Morphosyntactic variants are not always binary, however. In the expression of necessity/strong obligation in English, four variants can be used: *must*, *have to*, *have got to* and *got to*, as in (28–31).

(28) And we said, 'If you join the club, you *must* go to church.' (Tagliamonte and Smith, in press)
(29) And I *have to* wear a hearing-aid, 'cos I got tinnitus as well!
(30) You're told you'*ve got to* speak properly.
(31) You *got to* leave it up on t' hilltop.

Table 4.2 Overall distribution of variants of deontic modality

Variant	*No.*	*%*
must	62	10
have to	277	45
have/'ve/'s got to	214	35
got to/gotta	59	10
Total	612	100

Table 4.2 shows overall distribution of these forms across a range of dialects in the British Isles (Tagliamonte and Smith, to appear).

While Tables 4.1–2 show robust competition between forms, such is not always the case. This is demonstrated in the use of the *for to* infinitival construction, as in (32):

(32) He'd light a furnace *for to* wash the clothes. (Tagliamonte *et al.* in press)

Despite the prominence of this form in the history of English, our analysis of the same data set used for deontic modality (Table 4.2) showed that the varieties under investigation either had no occurrences of use at all or very few. Table 4.3 shows the overall distribution of use of the *for to* variant. Although there are many potential contexts of use of the *for to* infinitive (total contexts of use = 6,636), actual occurrences of the non-standard *for to* variant is miniscule (1.4 per cent). Such results are often indicative of an obsolescing feature: while in itself this is an extremely interesting finding, in reality there is little room for further analysis of forms – uncovering concurrence patterns or correlations is the next, and probably most revealing stage of the analysis.

Table 4.3 Overall distribution of *for to* infinitive

Word	*No.*	*%*
to	6,544	98.6
for to	92	1.4
Total	6,636	100

Revealing correlations

While overall distributions of forms indicate how common particular variants are, they shed little light on the processes underlying the choice mechanism. In order to do this, it is necessary to 'examine closely the forms that a linguistic variable takes, and note what features of the context co-occur with these forms' (Bayley 2002: 118). These include both surrounding linguistic environment as well as social

factors (see also Chapter 3). For example, consider examples (23–7) above. In these cases, there are two forms, *was* and *were*, but note that the features of the context in which they appear also vary: in (24) the subject type is second person singular *you*. In (25) it is a plural noun phrase. (27) is an existential construction. Moreover, (26) is uttered by a young female, whereas (25) is attributed to a young male. These different features of the context – both linguistic and non-linguistic – may influence whether a speaker chooses to say *was* or *were*.

In order to find out if this is indeed true, the coding system now becomes more elaborate – not only do we code for whether the variant is *was* or *were*, but we also code for the differing contexts of use or *factor groups*. The factor groups in this analysis are speaker information, subject type, polarity (whether affirmative or negative) and verb function. The data with contextual factors coded are shown in (23″ – 27″):

(23″) (Rc6AC) They *were* all in Gaelic.
(24″) (Sr2AC) *Was* you home?
(25″) (SanAA) The plans *was* drawn up.
(26″) (St4NA) We *wasna* actually gan thegither.
(27″) (RqtAC) There *were* four of us gied away with her to the blueberries.

The computer program 'reads' the data from left to right. In (23″) *R* signals the variant is *were*; *c* indicates that the utterance was spoken by an older male; *6*, that the subject type is third person pronoun *they*; *A* records that the **utterance** is affirmative; *C*, that the verb function is copular. In (26″) the variant is *was* (S), the speaker is a middle-aged male (t), the subject type is first person plural *we* (4), the utterance is negative (N) and the verb function is auxiliary (A). From this information the statistical program computes the various correlations and frequencies of use.

Table 4.4 provides the frequencies of non-standard *was* by one factor group – **age**. The oldest speakers use the highest rates of the non-standard form (58 per cent), the middle-aged speakers the lowest (35 per cent) and the young speakers (44 per cent) are situated somewhere in between. Table 4.5 shows the results for another contextual factor – grammatical person: there are high rates of non-standard *was* in all contexts except *they*, which is categorically standard.

Tables 4.4 and 4.5 demonstrate that there are correlations both with type of subject and with age: in other words, how many times *was* (or *were*) is used depends

Table 4.4 Overall distribution of *was* in *were* by age

	No.	*%*
Old	475	58
Middle	358	35
Young	518	44

Table 4.5 Distribution of *was* in *were* by subject type

Subject type	*No.*	*%*
Second singular *you*	161	69
First plural *we*	368	67
Second plural *you*	10	10
Third p. pronoun *they*	435	0
Existential *there*	162	90
NP plural	187	56
Relative pronoun	28	71

on the age of the speaker and what subject type is in the clause. Moreover, it is now easy to see why overall distributions only can often be 'deceiving' in that they hide more than they actually reveal. Table 4.1 showed that Buckie has 58 per cent non-standard *was*, which might lead us to expect this variant can occur anywhere. Table 4.5 demonstrates that this is not the case.

Let's now look further at deontic modality. Table 4.2 suggests that *got to*, as in (31), is used 10 per cent of the time in all dialects and there is a fairly even split between *have to* and *have got to*. But what happens when we divide the data into the different communities? Do they all pattern in the same way? Figure 4.1 shows the results. It shows that Tiverton is the only community which uses *got to* to any degree. In two communities (Cullybackey and Portavogie, and Buckie) the predominant form is *have to*, with much less use of *have got to*. Thus the communities are not equal with respect to the use of these four variants.

Once we begin to disentangle the correlations of these variants, we can see exactly *where* and *when* the variants occur. This allows us to go some way to explaining and interpreting the variation.

Uncovering competing influences

However, we still have one step further to go, as 'it is unlikely that any single contextual factor can explain the variability observed in natural language' (Bayley 2002: 118). The use of non-standard *was*, or zero relative, or copula deletion, or quotative *be like*, or indeed any other linguistic variable, is most likely the result of a combination of factors, whether age, speaker **sex**, subject type or polarity. Modelling this type of variation can be done by multivariate analysis, which can deal with these competing influences, as it permits us to model the combined contribution of all the contextual factors simultaneously. This type of analysis provides three important pieces of information: (1) which factor groups have a statistically significant effect on the choice of the particular variant (factor groups which are not significant are often shown in brackets), (2) which factor group has the strongest effect (shown by the largest *range*) and (3) which factors within the different factor groups favour (above 0.5) or disfavour (below 0.5) the variant.

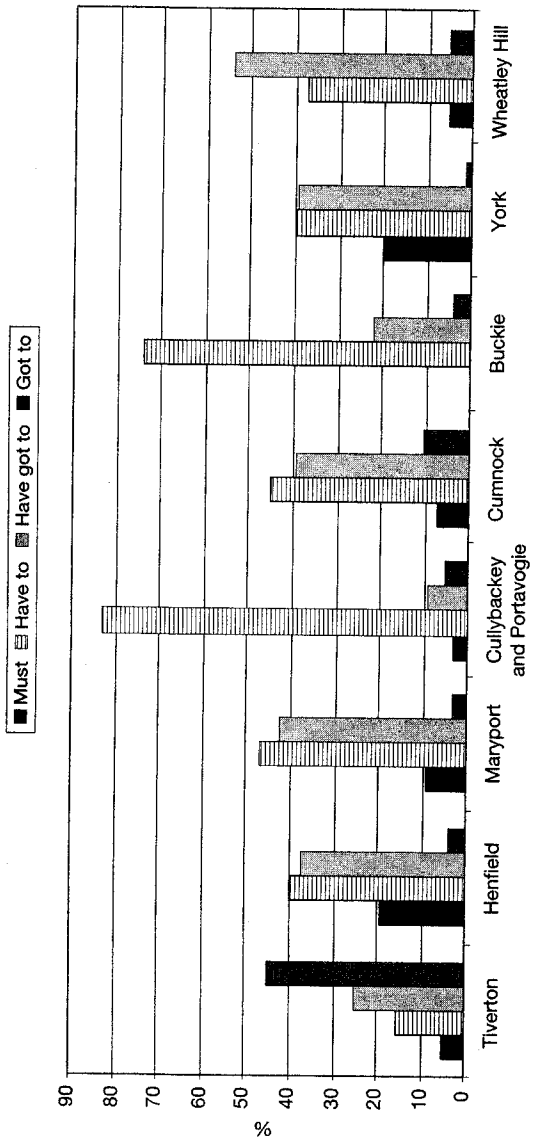

Figure 4.1 Use of forms for deontic modality across eight communities (Tagliamonte and Smith, in press)

Table 4.6 shows a multivariate analysis of the use of the probability of the *be like* quotative being used in the speech of young university students. It shows that all three contextual factors, speaker sex, grammatical person and the content of the quote (what is actually being reported), exert a statistically significant effect on the use of *be like*. The most significant factor group, that is, the one that exerts the strongest influence on the choice of *be like*, is speaker sex, with a range of 31. Moreover, it is favoured by females, in first person *I* contexts, when reporting some non-lexicalized sound, as in (33):

Table 4.6 Variable rule analysis of the contribution of speaker sex, grammatical person and content of the quote to the probability of the *be like* quotative

Speaker sex	
Female	0.67
Male	0.36
Range	31
Grammatical person	
First person	0.56
Third person	0.43
Range	13
Content of quote	
Direct speech	0.45
Internal dialogue	0.57
Non-lexicalized sound	0.67
Range	22

Source: adapted from Tagliamonte and Hudson (1999: 100–4).

(33) And I was like 'Whaaaam!' (Tagliamonte and Hudson 1999: 163)

Table 4.7 shows the results for *was/were* in the Buckie data, this time including percentages of number of contexts of use. (Note that third person pronoun *they* has been removed from the analysis, as it was categorically standard – Goldvarb deals with variable contexts only). As well as grammatical person and age, polarity (whether the sentence is positive or negative), verb type and the speaker's sex are also considered.

Table 4.7 shows that grammatical person and age are significant in the variation, while verb function, polarity and speaker sex do not exert a statistically significant effect on the variation (indicated by the brackets round the factor weights). In other words, if the speaker is older and using an existential construction, then they are likely to use non-standard *was*. If the speaker is middle-aged, on the other hand, and the subject type is full NP, then it is more likely that *were* will be used.

Table 4.7 Variable rule analysis of the contribution of factors to the probability of *was* in *were* contexts in Buckie, all speakers

Factor	*No.*	*Factor weight*	*%*
Grammatical person			
Second person singular *you*	161	0.49	69
First person plural *we*	368	0.44	67
Third person plural Full NP	187	0.33	56
Existential *there*	162	0.80	90
Range		47	
Polarity			
Affirmative	838	[0.50]	69
Negative	40	[0.56]	75
Function			
Copula	602	[0.48]	69
Auxiliary	276	[0.55]	69
Age			
Old	331	0.66	81
Middle	210	0.35	57
Young	337	0.44	65
Range		22	
Sex			
Male	438	[0.50]	71
Female	440	[0.50]	68
Total No.	878		

Note: Corrected mean 0.72.

Thus multivariate analysis allows us to view the *combination* of factors that influence the use of one form over another. For the case of quotative *be like*, it is speaker sex, what is being quoted and which grammatical person is used that all go into the 'mix' in the choice of *be like* over other quotatives. With non-standard *was*, age and grammatical person are the important influencing factors.

Conclusion

Utilising the Labovian paradigm, I have outlined some of the steps taken in the quantitative analysis of morphosyntactic variables in a range of dialects in the British Isles and elsewhere. I started with the initial steps of how to transcribe the data in order to ensure a consistent record of what was actually said. I then described what to exclude and include in the data, how to code the data ready for

statistical analysis and then how to model the multifaceted influences which are endemic in spoken data. Through these steps, the complex system of linguistic and social constraints on morphosyntactic variation can be uncovered.

FURTHER READING

Cheshire, J. (1982) *Variation in an English Dialect: A Sociolinguistic Study*, Cambridge: Cambridge University Press.

Milroy, L. and Gordon, M. (2003) *Sociolinguistics: Method and Interpretation*, Oxford: Blackwell.

Tagliamonte, S.A. (2006) *Analysing Sociolinguistic Variation*, Cambridge: Cambridge University Press.

Tagliamonte, S. and Hudson, R. (1999) 'Be like *et al.* beyond America: the quotative system in British and Canadian youth', *Journal of Sociolinguistics* 3 (2): 147–72.