Causal Interrogative Variation in Multicultural and Traditional Varieties of London English

James Brookes, David Hall, Jenny Cheshire and David Adger

Queen Mary, University of London

October 2017

Queen Mary's OPAL #36
Occasional Papers Advancing Linguistics

Causal Interrogative Variation in Multicultural and Traditional Varieties of London English

James Brookes, David Hall, Jenny Cheshire and David Adger
Queen Mary University of London

1 Introduction

The grammar of English has several interrogative expressions that serve to question an eventuality's reason or purpose, termed "causal interrogatives".[1] Alongside the basic *why*, as in 1a, we also find two other well-established variants—the discontinuous *what…for*, as in 1b, and *how come*, as in 1c.

(1) a. Why are you looking at me like that?
(Ryan; Sulema_Ryan_Kirsty; LIC)[2]

b. What you look at me like that for?
(Chris; Chris_Kim; LIC)

c. How come no one don't look here?
(Roshan; Roshan_Robert_Kevin_2; MLEC)

These three variants exist alongside each other in most, if not all, dialects and varieties of English.[3] Yet in one specific variety Multicultural London English

[1] In using "cause"/"causal" as a hypernym for both "reason" and "purpose", we follow Huddleston and Pullum (2002); for discussion, see below. *Why* can also be used as a relative, discussed below, but this usage is less basic; here we use "interrogative" as a fairly loose term that includes the relative case.

[2] The examples given here (and elsewhere, unless otherwise stated) are drawn from the two main corpora we make use of in this paper: LIC = *Linguistic Innovators*, MLEC = *Multicultural London English*. To reference the example, we first give the speaker's pseudonym, followed by the corpus file, and finally the corpus.

[3] There are several other, more peripheral ways, to question cause, e.g. *What happened to get you excluded? = Why did you get excluded?*, *What's the reason you did it now? = Why did you do it now?*, *What's the man doing in the garden? = Why is the man in the garden?*, etc. We leave such cases aside in the present study, as they are not as grammaticalized as the other variants we look at and, in any event, are likely to be very rare.

(henceforth: "MLE")— the emergent multiethnolect used by speakers in the multilingual environment of inner city London—a fourth structurally interesting variant of the pattern *why…for* has also been reported (e.g., Kerswill et al., 2004–2007, 2013), an example of which is given in 2.

(2) Why they looking at me like that for?
(Paul; Tacito_Paul_Tanya; MLEC)

Our concern in this paper is to explore and compare causal interrogative variation in Multicultural London English (MLE) and traditional London English. Using state-of-the-art data mining techniques, we endeavour to shed light on the factors that speakers of these two divergent London varieties draw on when making a choice for an interrogative expression. While our central focus is on *why…for*, we incorporate the other established variants into our analysis as well. Given that differences between MLE and traditional London English have previously been analysed as cases of sociolinguistic variation (e.g., Cheshire et al. 2011), we consider the potential influence of social factors in conditioning the use of the different interrogative forms. We find, however, that no social factor can fully explain the differences in frequency and distribution of use, and argue instead that a mixture of syntactico-semantic and psycholinguistic explanations are required to get at the full picture. From a cross-linguistic perspective, we show that morphologically distinct dedicated purpose interrogatives are not uncommon, and that *why…for* instantiates such a form in English. Furthermore, we argue that the relative high frequency of *why*, and the preference for short clauses with *why…for* can both be explained as instances of a related processing preference for *why*.

To provide context for our study, we begin by discussing prior findings on WHY-variants, and then outline the contents of our paper.

1.1 Prior Related Work

1.1.1 *Why*

Why would hardly appear to require any sort of introduction. It is a high frequency item, occurring over 1500 times per million words in spoken dialogue (BNC, demographic sample). And it is well established in the history of English—its present form derives from Old English *hwí/hwý*, a form which itself is the reflex of the Indo-European form **k^{w}ei*, the locative of the pronominal **k^{w}o-* 'who' (see *OED*, s.v. *why*). Despite its ubiquity, *why* deserves comment as it exhibits some interesting syntactic and functional characteristics, a discussion of which will provide a useful backdrop for the discussion of the other variants of interest.

In terms of its semantics, *why* is strongly polysemous: it can question both reason and purpose aspects of cause (Huddleston and Pullum, 2002, 725–726, 906; cf. Quirk et al., 1985, 564, 695), as the following example demonstrates (adapted from Huddleston and Pullum, 2002, 726):

(3) a. [A:] Why did you get up so early?

b. [B:]

i. Because I couldn't sleep. [REASON]

ii. To do some gardening while it was still cool. [PURPOSE]

Unlike the other causal interrogative variants (at least, *what…for* and *how come*), *why* has various syntactic and functional peculiarities. In addition to its canonical illocutionary force of a question, *why* can also function as a directive, either with positive/negative bare infinitivals or in a negative finite clause (Huddleston and Pullum, 2002, 835, 874, 941), e.g.

(4) a. Why give her anything (*something)? [POSITIVE INFINITIVAL]

b. Why not give her something (*anything)? [NEGATIVE INFINITIVAL]

c. Why don't you give her something (*anything)? [NEGATIVE FINITE]

As should be clear from these examples, in these contexts *why* behaves somewhat strangely with respect to structural patterning and polarity effects. First, bare infinitivals are possible. In general, bare infinitivals are not possible in interrogative clauses. Second, a positive clause would typically trigger the positive polarity item *something*, whereas a negative clause would trigger the negative polarity item *anything*. However, here the reverse is the case, which is due to the clause's semantics: the positive clause triggers the negative polarity item *anything*, and the negative clause triggers the positive polarity item *something* (Huddleston and Pullum, 2002, 835; Quirk et al., 1985, 820–821, 840). This directive use of *why* seems impossible with *what…for* and *how come* (no evidence for *why…for* though):

(5) a.

i. *What give her anything for?

ii. *What not give her something for?

iii. *What don't you give her something for?

b.

i. *How come give her anything?

ii. *How come not give her something?

iii.??How come you don't give her something?

In addition to its use as an interrogative, *why* can also be used in relative constructions in which the antecedent is *reason* (Huddleston and Pullum, 2002, 1051). However, *what…for* cannot be so used, even if its form modified to a standard relative; and neither can *how come*.

(6) a. The reason why he did that was…

b.
   i. *The reason what he did that for was…
   ii. *The reason that/which/∅ he did that for was…

c. *The reason how come he did that was…

In terms of its grammatical function in the clause, *why* is typically used as an adjunct, as in all the examples given thus far, but it can also be used as complement of *be*—for instance, in *it*-cleft constructions (Huddleston and Pullum, 2002, 906). Such contexts are not possible for *what…for/how come*, or at least have questionable acceptability status.

(7) a. Why is it that these views exist?

b. *What is it that these views exist for?

c. *How come it is that these views exist?

Finally, while all the three established variants can be used in contexts of complete clausal ellipsis 8, only *why* can be used when only nominal material remains 9.

(8) a. [A:] Mary has to go to the hospital later this afternoon.

b. [B:]
   i. Why?
   ii. What for?
   iii. How come?

(9) a. Why no classes today? (Quirk et al., 1985, 840)

b. *What no classes today for?

c. *How come no classes today?

1.1.2 *What…for*

The variant *what…for* has been variously characterized as "colloquial" (*OED*; Claridge, 2012, 177, 193), "idiomatic" (Huddleston and Pullum, 2002, 906; Claridge, 2012, 177), "informal" (Huddleston and Pullum, 2002, 906; Claridge, 2012, 177), and "polite" (*OED*).

It is first attested in its modern form at the end of the 16th century, with precursor structures dating back to 1200 (Claridge, 2012, 189), but there is some disagreement as to its exact origin. Zwicky and Zwicky (1971) surmise that the "obvious source" of the variant is ultimately based on the prepositional phrase *for what purpose*, with deletion of the nominal head *purpose* to give *for what*, and subsequent fronting of *what*. Claridge (2012) shows via empirical evidence that this cannot be correct. *For what purpose* (or *for what* followed by semantically similar nouns such as *end, reason, cause*) is attested much later than *for what* without a following head noun. Further, all the nouns are loan words which are attested later than the bare *for what*. Thus, Claridge argues, *for what* must be the origin of the modern day interrogative structure *what…for*. She proposes an origin which is based on the empirical facts, with *for what* directly giving rise to *what…for* in Middle English, when preposition stranding took root. Below, we will see that *why…for* may have undergone a similar diachronic evolution.

In terms of its semantics, it has been assumed that *what…for* questions only the purpose aspect of cause (Zwicky and Zwicky, 1971). Thus, while 10 is apparently semantically OK, 11 isn't.

(10) What did you hit the child for?

(11)??What were you ill for?

However, in a large corpora of historical data, Claridge (2012) finds primarily purposal, primarily causal and ambiguous examples, which may fit more comfortably with the semantic range of *for* as outlined by Bresnan (1972, 79–81). Thus, it is not clear whether Zwicky and Zwicky (1971)'s assertion is correct.

1.1.3 *How come*

Like *what…for*, *how come* has similarly been characterized as "informal' (Claridge, 2012, 177; Huddleston and Pullum, 2002, 909; Quirk et al., 1985, 840), "colloquial" (Claridge, 2012, 177), and "idiomatic" (Claridge, 2012, 177; Huddleston and Pullum, 2002, 908–909), although its first attestation comes much later, in the 18th century (Claridge, 2012).

Structurally, it is quite different from the other variants considered here as *how come* typically, although not always, functions as a fossilized multi-word unit. First, the *come* part is not conceived of as a verb, as it receives no inflection, behaving much like a particle. Second, *how come* matrix clauses do not exhibit subject-auxiliary inversion, unlike other interrogative matrix clauses. And third, *how come* has scope only over the immediate clause. Thus, in 12, *how come* questions *you thought*, while *why* (in 13) and *what…for* (in 14) are ambiguous: they could question the reason/purpose of John's going to London or your belief of the fact.

(12) How come you thought John went to London?

(13) Why did you think John went to London?

(14) What did you think John went to London for?

Semantically, *how come* has been argued to focus on the reason aspect of cause (Zwicky and Zwicky, 1971).

1.1.4 *Why…for*

Why…for, the variant of central interest in the present work due to its apparent MLE roots, has received hardly any scholarly attention, with the exception of a few remarks scattered about in the literature on grammatical variation in MLE (e.g. Kerswill et al., 2004–2007, 2013; Cheshire et al., 2013, 2015; Coveney and Dekhissi, 2017).

Syntactically, it appears to be a structural hybrid of *why* and *what…for* (Kerswill et al., 2004–2007; Coveney and Dekhissi, 2017). The formation is somewhat comparable to several other multi-word interrogative expressions, such as *where…to* and *how long…for*, in which the final preposition is semantically redundant. This redundancy has been seen as a "reinforcing" or "expressive" feature of the *why…for* question frame (Coveney and Dekhissi, 2017).

Perhaps in view of that, the prior literature has pointed out that the variant typically occurs in pragmatically-charged environments, such as those that involve a confrontation between the participants in the conversational exchange or in direct speech contexts that report an aggressive encounter. This can be seen in the following example, where the speaker recalls an occasion when he and his friends were mugged.

(15) and then ... like they were pulling him about he was stopping them from taking stuff and everything but they managed to get his phone . and his house keys but then he said "why are you taking my house keys for?" . and he dashed his house keys at him
(Dean; Dean_Chris; LIC)

In terms of its semantic distribution, Coveney and Dekhissi (2017) assume that *why…for* is essentially a purpose-questioning interrogative rather than one that questions cause in general, as we have already noted Zwicky and Zwicky (1971) argued for *what…for*. As Coveney and Dekhissi (2017) note, *why…for* typically occurs in clauses with verb phrases headed by eventive verbs that project agentive first arguments (e.g. *ask, do, lie, speak, try*) rather than those headed by stative verbs (e.g. *fear, be happy*) or by eventive verbs with thematic subject arguments (e.g. *die*).

Coveney and Dekhissi (2017) enumerate a number of apparent syntactic restrictions on *why…for*'s usage. For instance, the pattern is apparently not found in elliptical structures or in negated clauses. They also state that the structure is rare in embedded interrogatives, implying that the pattern emerged as matrix-only phenomenon interrogatives and then eventually spread into embedded. However, it could simply be there are far few tokens in their dataset to make any reasonable assumption about these apparent restrictions.

1.2 Structure of Paper

The rest of this paper in structured as follows. In Section 2, we detail the London corpora we used in our study, and Section 3 discusses how we extracted the relevant data and its cleaning. Section 4 presents the features that we use in our model of WHY-variation. In Section 5, we discuss our machine learning approach, and present the results. We discuss and interpret our findings in Section 6. Section 7 offers some brief remarks on the history and evolution of *why…for*, and Section 8 concludes.

2 Corpora

The primary data used in this paper are drawn from two existing spoken corpora of Multicultural London English, the result of two recent research projects—*Linguistic Innovators: the English of adolescents in London* (Kerswill et al., 2004–2007) and *Multicultural London English: the emergence, acquisition and diffusion of a new*

variety (Kerswill et al., 2007–2010). We present a brief overview of the original projects here and, in doing so, detail the sub-corpora we focus on specifically.

The *Linguistic Innovators* project was concerned with describing innovative features of London English and exploring the geographic and social drivers behind them, by comparing the speech of inhabitants in two different London locations—the inner city borough of Hackney in the east of London and the outer London borough of Havering further to the east (see Figure 1). These locations were selected because of their starkly different socio-demographic characteristics. For instance, Hackney is the third most ethnically diverse borough in London, beaten only by Newham and Brent, while Havering, together with neighbouring Bexley and Bromley to the south, is the least. This is reflected by the following statistics: Hackney has a much lower percentage of Anglo[4] inhabitants than Havering (44.1% vs. 92.0%), a higher rate of ethnicity mixing within households (35.6% vs. 8.0%) and of immigration/in-migration (7.6% vs. 3.9%). However, it is less affluent compared with Havering, on established indicators of affluence: e.g. car ownership (44.0% vs. 76.7%), home ownership (32.1% vs. 79.7%), household overcrowding (27.6% vs. 5.6%), and D/E social grades (39% vs. 31%).[5]

These striking social differentiations allowed the project to research differences between the two locations. To capture the language of mid-2000s London English, adolescents in post-16 education were recruited in the two locations, on the basis that they are more likely to use innovative features than other age groups. Recordings took place mostly in the colleges in each location in pairs, groups or individually, and usually in the presence of a fieldworker. Some of the recordings were done individually. The Hackney sub-corpus encompasses 50 recordings of 49 16–19 year olds. 27 of these speakers are male and 22 are female, and 22 are Anglo and 27 are Non-Anglo. The Havering sub-corpus encompasses 46 recordings of 49 16–19 year olds. 27 of these speakers are male and 22 are female, and 36 are Anglo and 13 are Non-Anglo. Also interviewed in the *Linguistics Innovators* project were a handful of elderly speakers born between 1918 and 1938—however, as these were all Anglos and as few speakers were interviewed ($n = 6$), we do not include them in this study in order not to bias the results.[6] This Hackney sub-corpus of *Linguis-*

[4]Following previous work on MLE, we use the term "Anglos" defined as "children of families with more than three generations' settlement history in the fieldwork area" (basically White British in the census data) and "Non-Anglos" as "children/grandchildren of immigrants with different ethnic backgrounds but representative of the ethnic distribution of the fieldwork area" (Kerswill et al., 2007–2010, 3).

[5]Note that these figures relate to the 2001 census (the last census to take place before *Linguistic Innovators* began) rather than the most recent 2011 census. Data drawn from `http://www.nomisweb.co.uk/census/2001`.

[6]We should note that these elderly speakers do not contribute any *why...for* observations.

tic Innovators contains approximately $800,000$ tokens and the Havering sub-corpus about $700,000$ tokens.[7]

While the *Linguistics Innovators* project focused on innovations, the *Multicultural London English* project had as its overarching goal to explore how MLE arose. This latter project focused exclusively on the language in the multiethnic centre of inner city London, sampling again from the borough of Hackney, as in the earlier project, but also neighbouring Islington and Haringey, which are similarly ethnically diverse (see Figure 1). In addition to interviewing 16–19 year olds, this project covered a number of other age groups: children at various stages of the L1 acquisition of MLE (4–5, 7–8, 11–12 year olds), younger adults aged 25–30, and caregivers aged around 40 (many of whom were L2 speakers of English). In total, 127 speakers were recorded—67 female and 60 male, and 95 Non-Anglo and 32 Anglo. However, in the present study, we focus exclusively on the 16–19 year olds, in order to ensure age-group comparability with the *Linguistic Innovators* Hackney and Havering sub-corpora.

[7] Raw text as tokenized by NLTK using `RegexpTokenizer(r'\w+')`.

Figure 1: Map of London showing the boroughs studied in the *Linguistic Innovators* and *Multicultural London English* projects

3 Data Extraction

To extract causal interrogative observations from the above-mentioned corpora, we used Python, the text analytics package **NTLK** (Bird et al., 2009), and the data manipulation library **pandas** (McKinney, 2010). A program was written to automatically grab from each transcript all word tokens of *why*, *what* (if a *for* followed up to 40 tokens downstream), and *how* (if *come(s)* followed immediately). In addition, for each token we extracted a context window of ± 150 words surrounding the interrogative token, the speaker's ID,[8] as well as file metadata.

[8]This was identifiable for the most part by extracting the token immediately before the last ':' in the prior context. In some cases, e.g. due to an intervening short turn, erroneous speaker labels were

Following the initial collection procedure, we manually read through each observation's context, identifying and removing false positives. For the most part such cases involved *what*. These included examples in which the captured *for* did not participate in any type of dependency relation with *what*. We also removed cases in which the *what…for* was part of a dependency structure, but was unambiguously not equivalent to *why*, as in the examples in 16, where *what* replaces an NP.

(16) a. that's not **what** Hackney people are looking **for**
≠ that's not why Hackney people are looking

b. **what** do you take me **for**?
≠ why do you take me?

In some cases, *what…for* sentences were fully ambiguous between *why*-alternants and NP replacements, as in 17.

(17) my mum was ready to beat me up like when I was in there [prison] but she didn't know **what** I was in there **for**
= what (e.g. crime) I was in there for
= why I was in there

We felt it best to retain such examples in the analysis, for the simple reason that if *why…for* had occurred in its place with the exact same clausal material, we would almost certainly have wanted to include it.

Observations exhibiting false starts, unfinished clauses, and other unparsables were excluded from the analysis.

We removed from the dataset all observations from speakers for whom social profiles were not available, i.e. the fieldworkers, tutors, and other individuals. It is difficult to know what to do with deterministic individuals, i.e. speakers who use only one of the variants, usually the most common one. Tagliamonte and Baayen (2012, 165) note that variationist studies either exclude non-variable speakers altogether or include them "on the assumption that internal predictors will be parallel across individuals". In this research, we include such speakers.

It is standard practice in variationist studies to exclude so-called "categorical contexts"—i.e., those contexts where the target class is perfectly predictable (see e.g. Tagliamonte, 2006, 86ff.). For instance, in our study, directive causal interrogatives constructed with *not* and a bare verb form (e.g. "Why not do that?") are always realized by *why*, and thus *why* is perfectly predictable. In the present study we do not follow this variationist practice for three central reasons. First, categorical contexts are part of one's linguistic knowledge, and need to be explained and

captured. These were manually identified and corrected.

modelled. Second, as our dataset is quite small, it is unclear whether an apparent categorical context would be truly categorical if more data were available. And third, traditional statistical models such as classical logistic regression which are usually used in variationist research cannot handle cases where the data are perfectly or partially separable on a particular feature; in our research, we make use of machine learning models where this is no longer poses a problem.

Table 1 presents a summary of the data. We note the severe class imbalance in that the relative frequency of *why* is much greater than those of the other three variants, and the absolute frequencies of *how come* and *what…for*, in particular, are much too low to enable traditional regression modelling with a relatively large feature set. We will present a modelling strategy below that overcomes these problems.

	Linguistic Innovators Corpus		Multicultural London English Corpus	
	Outer London	Inner London	Inner London	Total
how come	5 (1.69%)	12 (2.69%)	9 (3.72%)	26
what…for	9 (3.05%)	16 (3.59 %)	11 (4.55%)	36
why	258 (87.46%)	384 (86.10 %)	202 (83.47%)	844
why…for	23 (7.80%)	34 (7.62 %)	20 (8.26%)	77
Total	295	446	242	983

Table 1: Summary of WHY data

4 Features

In this subsection, we outline the features we annotated the dataset for. As is becoming usual in corpus linguistics and variationist research, we include (i) sociodemographic[9] attributes of the speaker that contributed the observation; (ii) temporal features to capture diachronic variation; (iii) grammatical attributes pertaining to the clause the observation was extracted from; (iv) general psycholinguistic and production processing attributes; and (v) attributes pertaining to the local conversational and narrative context.[10] Given the dearth of research on WHY-variation, it is unclear what might control it. Thus, in accordance with the data mining approach we pursue, our annotation coverage is extensive and somewhat exploratory.

[9]Attributes for the speaker's identity, the conversation, and the speaker's self-defined ethnicity were also annotated, but we do not include these in the models because of the vast numbers of levels rendering at least one of our machine learning model analyses (random forests) computationally intractible.

[10]This is defined as a window of ± 150 words surrounding the variable site.

4.1 Sociodemographic Features

Ethnicity: A speaker's ethnicity is an obvious candidate feature when considering linguistic variation in a multiethnic population such as that in London. Prior research in the MLE space has demonstrated that ethnicity is an important discriminator at various levels of linguistic analysis, including grammatical variation. For instance, in the *a/an* + vowel alternation, Gabrielatos et al. (2010) found that speakers with non-Anglo backgrounds are three times more likely to use the non-standard indefinite *a* + vowel structure (e.g. *a apple*) compared with Anglo speakers. In addition, Cheshire et al. (2013) explore the innovative use of *who* as a topical marker and find that this usage is driven by non-Anglo speakers. Following previous MLE research (e.g. Gabrielatos et al., 2010), this variable was categorized using two feature levels: *Anglo* for those speakers who report their ethnicity as "White British" and *Non-Anglo* for all other speakers. Given the above, if *why…for* is an innovative structure, we might expect the non-Anglos to be leading in its usage.

Sex: There is some evidence that innovative linguistic forms in MLE are more likely to be adopted by males. For instance, in the domain of phonology, males exhibit more "extreme" diphthong realizations (Cheshire et al., 2011); in grammar, Cheshire (2013) finds that the use of *man* as a pronoun is used by males in her corpus with only two exceptions; and, *a* + vowel is used by males almost twice as much as females (Gabrielatos et al., 2010). If *why…for* patterns like other MLE innovations, we might expect males to be leading its usage.

Residence: As pointed out above, inner city London offers a rich environment for the formation of new grammatical structures because of its multiethnic and multilingual make-up, in contrast to outer of London (Cheshire et al., 2011; cf. Cheshire et al., 2015). Previous MLE research that has examined linguistic variation has found that inner city London leads on several innovation variables — for instance, *who* has a topic-marking strategy (Cheshire et al., 2013) and indefinite *a* + vowel (Gabrielatos et al., 2010). On this basis, we might expect *why…for* to be relatively more frequent in inner city London, assuming it is an innovation. Our annotation scheme for this feature has three levels—*Inner London* for those speakers who live and whose conversations were recorded in the multicultural boroughs of Hackney, Haringey, and Islington; *Havering* for those speakers who live in the outer London borough; and *Commuter to Havering* for those speakers who commuted to Havering from other parts of London.

Friendship Network: In the *Linguistic Innovators* project, the speakers were given a score of 1 through 5 relating to the ethnicity profile of their friendship

network, with a score of 1 indicating that all friends had the same ethnicity as the speaker and a score of 5 indicating that more than 60% of friends had a different ethnicity than the speaker. In Gabrielatos et al. (2010)'s *a/an* alternation study, it was found that users of *a* + vowel had on average higher friendship network scores than non-users, an effect that is somewhat stronger for Anglo speakers. We surmise, then, that a speaker's friendship network score might be a relevant discriminator in WHY-variation, with speakers with higher friendship network scores using the *why…for* variant relatively more frequently than those with lower scores. To reduce data sparsity, we collapsed the 1–3 categories together as *Low* and the 4–5 categories as *High*. As information on friendship networks was not recorded for speakers who participated in the subsequent *Multicultural London English* project, we annotated observations from this latter corpus with an explicit *Unrecorded* tag.

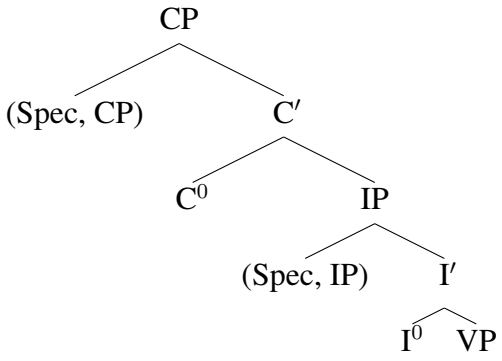
4.2 Temporal Features

Project: In order to gauge possible short-term temporal changes in WHY-variation we include an indicator for the project—*Linguistic Innovators* or *Multicultural London English*.

4.3 Linguistic Features of the Clause

Using the generative-inspired clausal architecture in 18, our grammatical features naturally fall into the following groups: (i) properties of the CP layer; (ii) properties of the IP layer; and (iii) properties of the VP layer.

(18)



Properties of the CP: We included a feature for **Clause Type** (*matrix interrogative, embedded interrogative, relative*).

Properties of the IP: We included features relating to the **Subject Type** (*first person, second person, third person, elided, unmarked*); **Tense** (*present, past, untensed, elided*); **Progressive Aspect** (*progressive, nonprogressive*); **Presence of Modal Auxiliary** (*present, absent*); and **Polarity** (*positive, negative*).

Properties of the VP: To capture the distinction between purpose and reason interpretations in causal interrogative clauses, we annotated for whether the lexical verb's subject argument was volitional or non-volitional (**Volitionality**).

4.4 Psycholinguistic Features

Intervening Distance: We measured the length of the clause (and all its dependent clauses, if present). This variable serves as a proxy for the distance over which *for* has to be retained in working memory. We used a normalized version of intervening distance, squeezing the variable between 0–1 so it's within the same range as the dummy codes used for the categorical variables.

Lexical Persistence: It has been shown that speakers typically re-use lexical material they have encountered before ("beta-persistence" in Szmrecsanyi (2006)'s parlance). We decided to include binary variables for whether each of the lexical elements that make up the alteration (i.e., *why, what, for, how, come*) appeared in the 150 words preceding the variable site.

4.5 Features of the Conversational Context

We next annotate for features relating to structural and stylistic aspects of the conversational context of the WHY utterance.

Direct Speech: In previous research, it was suggested that *why…for* may be relatively more frequent in direct reported speech. Thus we include a feature for whether the WHY variant occurs in *direct* reported speech or in *regular* conversational discourse.

Turn Position: We coded a binary feature relating to whether the interrogative clause was placed in initial position in the turn (or, if in direct reported speech, in initial position in the imitated turn) or elsewhere.

Confrontational Index: As noted above, prior research has suggested that *why…for* may occur relatively more frequently in in confrontational or argumentative contexts (Cheshire et al., 2015, 15). In order to annotate this feature programmatically, we made various assumptions about the nature of confrontational discourse. Drawing inspiration from Hutchby (2013), we assumed that confrontational discourse may be more likely to contain (i) second-person directed threats (e.g. *I'll staple gun your bum*); (ii) second-person directed insults (e.g. *you little bitch*); (iii) "hostile" imperatives requesting that (an)other participant(s) in the exchange leave (e.g. *fuck off*) or be silent (e.g. *shut up*). For each observation, we calculated the following:

Number of threats To establish what might constitute a threatening utterance in our corpus, we first extracted from the entire corpus all unique strings containing a first person singular pronominal subject combined with a futurate marker (i.e. *I'll, I will, I'm gonna, I am gonna, I'm going to, I am going to*) followed up to four words downstream by any second person pronominal (*you, your, yours, yourself, yourselves*).[11] The resulting strings were then manually sifted to filter out obviously non-threatening utterances (e.g. *I'll come with you*). As a result, we retained 28 threatening utterance types (e.g. *I'll kill you*).[12] With these established, we subsequently used regular expressions to count their string frequency within the ± 150-word context of the observation.

Number of insults To establish a lexicon of insulting words, we first extracted from the corpus all unique strings starting with *you* and ending with a noun (up to 4 words away). We manually went through each of these strings, checked whether they were potentially offensive, and then counted their occurrence in the context.[13]

Number of hostile imperatives For this component, we counted the frequency of the following imperative verb sequences within each observation's context: *get lost, get away, go away, fuck off, piss off, back off, bugger off, fuck off, piss*

[11] We focus on threats made by a speaker to his/her interlocutor, as these are likely to be the most directly threatening.

[12] The verbs involved were as follows: *arrest, bang, bash, be rude with, beat, bite, break, bully, chase, disowe, end, go mad at, head butt, hit, hook, hurt, kick, kill, knock, punch, push, shank, shoot, slap, smack, smash, stab, staple gun.*

[13] The full list is: *you bastard, you battyboy, you batty boy, you bitch, you cock head, you crackhead, you cunt, you dickhead, you faggot, you freak, you goon, you idiot, you kiddie fiddler, you knob, you liar, you loser, you mammal, you moron, you muppet, you nonce, you nutter, you paedophile, you paki, you pervert, you piece of shit, you prick, you queer, you retard, you sado, you scumbag, you slag, you slob, you slut, you sod, you tart, you tramp, you twat, you wanker, you waste gash, you waste man, you waste men, you wasteman, you wastemen, you whore, you wimp.*

off, sod off, leave me alone, leave us alone, shut up.

To provide a single index of confrontational diction, we summed these values, and then [0,1]-scaled to bring the values in line with our binary and dummy-coded categorical variables.

Involvement Index Given that confrontational discourse often occurs in highly animated, energized, interactive, and dynamic contexts (see e.g., Hutchby 2013 for discussion), it could be that it is not the confrontational aspect of the encounter per se that makes *why…for* more likely, but rather that *why…for* is typically used in vivid and lively styles, which confrontational styles belong to. We therefore include a feature for how 'involved' the style was surrounding an observational unit.[14] We defined this feature to have the following sub-components:

Number of second person pronominals For each observation, we counted the number of second person pronouns in the discourse context (as stated earlier, this is 150 words either side of the interrogative).

Number of overlaps Overlaps are another indicator of interactive conversational styles. In the transcription, a single slash ('/') is placed at each end of the overlap for both speakers—thus, 4 slashes correspond to a single overlapping event. We used a regular expression to count the number of slashes in the extracted ± 150-word context for each observation, and divided these counts by 4.

Number of turns Assuming a higher turn-exchange rate is indicative of an interactive and involved style, we give a measure for the turn-exchange rate in the context, we used a regular expression to count the number of colons, used by the transcribers to indicate turn initiations.

Number of direct speech initiations The above variable does not take into account the turn structure within direct speech, which may be important (for instance, when a speaker recalls a confrontational encounter using direct speech). To account for it in our models, we included a variable for the number of times direct speech is initiated in the ± 150-word context. We used a regular expression to count the occurrences of double graves (``), signalling the onset of direct speech in our tokenized file.

[14] We are using 'involvement' in the sense of e.g. Biber (1991).

Again, to provide a single measure, we summed together scaled versions of these indices, and further scaled.

4.6 Summary

Table 2 lists the features we utilize in our study.

Feature Domain	Feature	Levels/Values
Sociodemographic	Ethnicity	Anglo, Non-Anglo
	Sex	Female, Male
	Residence	Inner London, Havering, Commuter
	Friendship Network	Low, High, Unrecorded
Temporal	Project	Linguistic Innovators, Multicultural London English
Grammatical	Clause Type	Matrix, Embedded, Relative
	Subject Type	1st/2nd/3rd person, Ellipted, Unmarked
	Tense	Present, Past, Untensed, Ellipted
	Progressive Aspect	Progressive, Non-progressive
	Modal Auxiliary	Present, Absent
	Negation	Positive, Negative
	Volitionality	Volitional, Non-volitional
Psycholinguistic	Intervening Material	$[0, 1]$
	Lexical Persistence	Lexeme Present, Absent
Conversational Context	Direct Speech	Direct, Regular
	Turn Position	Initial, Elsewhere
	Confrontation Index	$[0, 1]$
	Involvement Index	$[0, 1]$

Table 2: Summary of Features

5 Machine Learning Models and Results

5.1 Classification Models

Our task is determine which of the variants of WHY-variation, {*why, why…for, what…for, how come*}, a speaker of London English will use based on the features listed in Table 2. In doing so, our aim is to understand (i) how distinguishable the variants are from each other, (ii) which features are relevant for distinguishing variants, and (iii) the direction of each relevant feature's effect.

A wide range of classifiers is available to the variationist to model this kind of data, including, but not limited to, generalized linear models with logit link

function ("logistic regression") (e.g. McCullagh and Nelder, 1983), support vector machines (e.g. Vapnik, 2000), and random forests (Breiman, 2001). Logistic regression models, and their hierarchical extensions, are perhaps the most widely used within variationist research in linguistics. To cite just a few examples, Bresnan et al. (2007) use logistic regression to predict the dative alternation, Hilpert (2008) uses it to interrogate comparative choice, and MacKenzie (2013) illuminates predictors of auxiliary contraction by using such models. Tree models and random forests have also gained in popularity, particularly since the publication of Tagliamonte and Baayen (2012)'s inspiring paper. However, certain statistical learning techniques have not been used at all in language variation classification tasks, at least to our knowledge—e.g., penalized logistic regression, k-nearest neighbors, and support vector machines, although they are used in tasks in related fields such as computational linguistics and natural language processing.

Given that different classification models perform differently in different circumstances, we trained six different types of classifiers on our data: (i) penalized logistic regression (lasso), (ii) k-nearest neighbors, (iii) classification trees, (iv) bootstrap aggregation of trees (bagging), (v) random forests, and (vi) support vector machines. For details about these algorithms, we refer the reader to Hastie et al. (2001).

In terms of implementation, although the models we use do have multi-class extensions, we used one-versus-one classifiers, as binary contrasts are generally easier to interpret. That is, we trained 6 classifiers for each classifier type, and average model accuracies obtained through leave-out-one cross-validation.

As we have already noted, the datasets are characterized by extreme class imbalance: for instance, *why* makes up more than 80% of observations overall. For machine learning models, this can adversely affect predictive accuracy and interpretation. To address this issue, we downsampled the majority class for each classifier. To exemplify, for the *why/why…for* contrast, we randomly sample $n = 77$ *why* observations to match the $n = 77$ *why…for* observations. However, due to the downsampling technique, different results can obtain from different cross-validation runs, so we repeated the procedure 100 times and averaged results.

Table 3 presents average model accuracies by model and by variant.

We see that in general, the random forest classifier works the best (mean accuracy: 75.48%), followed closely by the lasso regression (mean accuracy = 74.57%). The worst performing classifier is the k-nearest neighbors, with a mean accuracy of 69.76%. Looking at which variants, we see that *what for* vs. *how come* is the most distinguishable pair (83.05%). This might be expected, given that the prior literature has drawn a sharp contrast between the two in terms of their functions (e.g., Zwicky and Zwicky, 1971). The least distinguishable variants are *why…for* vs. *what…for*. Indeed, the models for this contrast exhibit some anti-prediction,

	Classification Contrast						
Model Type	*why* vs. *why…for*	*why* vs. *what…for*	*why* vs. *how come*	*why…for* vs. *what…for*	*why…for* vs. *how come*	*what…for* vs. *how come*	Mean
Lasso Logistic Regression	81.47	71.93	76.02	49.01	81.19	87.79	74.57
k-Nearest Neighbors	76.60	69.53	72.44	47.53	74.60	77.88	69.76
Classification Tree	77.16	64.93	69.02	52.10	75.29	82.44	70.16
Bagging	80.53	73.96	75.65	50.78	78.81	80.92	73.44
Random Forests	81.40	75.39	75.37	48.46	82.87	89.42	75.48
Support Vector Machine	80.79	71.39	72.17	40.65	80.69	79.87	70.93
Mean	79.66	71.19	73.45	48.09	78.91	83.05	

Table 3: Classification Results (% classified correctly)

which we take to indicate that they are behaving spuriously. This result is to be expected, however, given the variants' very close structural similarities (see e.g., Coveney and Dekhissi, 2017, who note that in some cases *why* and *what* may be perceptually very similar in terms of their phonetics).

Due to the overall superiority of the random forest classifier, we discuss the results of this model in more detail in what follows. However, given that *why…for* vs. *what…for* discrimination is quite poor, we drop all further discussion of this part of the alternation.

5.2 Feature Importances and Directionalities

Determining how separable the variants are is only part of the story. It is also crucial to understand which (hopefully small) subset of the features drive separation, and how. Unfortunately, machine learning models do not readily lend themselves to interpretation as classical linear models do, so we approach inference indirectly as follows.

First, to determine the influence of a given feature, we used an inbuilt variable importance algorithm for the chosen random forest classifier. Simplifying somewhat, the variable importance of a given feature is the decrease in accuracy that results from randomly permuting that feature's values compared with a model containing the original feature values. Relevant features are those which cause the model to exhibit large decreases in accuracy above a threshold θ when their values are permuted.[15] So as to identify the most reliable predictors of the alternation, we regarded only those features whose importance score exceeded θ in 95 out of 100 forest iterations as being fully worthy of interpretation.

Second, in order to explore how each feature contributes to the model fit, we computed partial dependency information. Essentially, this means predicting the

[15] θ is defined to be the absolute value of the least influential feature's importance score; see Strobl et al. (2009) for the rationale behind this rule of thumb.

probability of a class label (e.g. *why…for*) for each value of a feature of interest while holding other relevant features at a default value (i.e. mode or median values).

5.2.1 *Why* vs. *why…for*

Figure 2 gives variable importances for the *why/why…for* model. Features shaded in black indicate a feature worthy of interpretation, with the number aside the bar giving the proportion of times that feature's importance score exceeded the relevance threshold θ.

The most important feature for distinguishing *why* from *why…for* is (1) clause type, which is followed by (2) subject type, (3) progressive aspect, (4) polarity, (5) direct speech, (6) turn position, (7) tense, and (8) intervening distance. All of these features relate to the interrogative's grammatical structure, psycholinguistic aspects, or conversational context. None of the social-demographic features we included—such as residence, network score, or ethnicity—appear to exert influence. The alternation is thus largely linguistically governed.

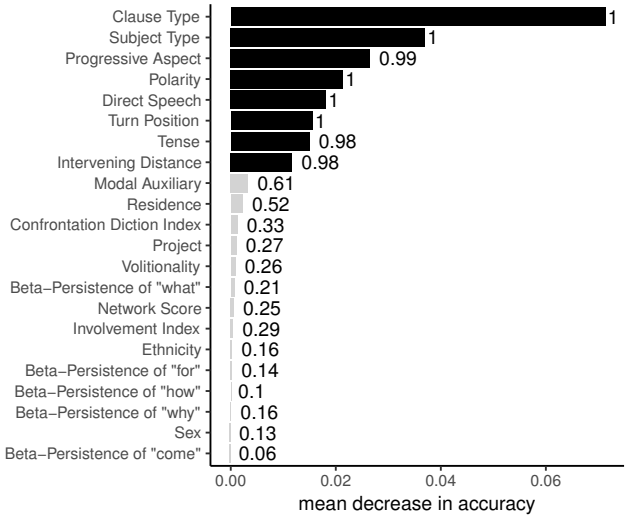


Figure 2: Variable importances for features in the *why/why…for* model. Features with bars shaded in black are deemed to be of relevance. The values to the right of each bar give the proportion of times the feature was deemed to be relevant out of 100 forest iterations.

Turning to effect directions, Table 4 shows the predicted probability of *why…for* for different levels of a given categorical feature while holding the other influential variables at their default values. This shows that, compared with *why*, *why…for* is

relatively more likely to occur in matrix clauses, with a second person subject, progressive aspect, positive polarity, and present or omitted tense marking. In addition, *why…for* is relatively more likely to occur in direct speech contexts than bare *why*, and in an initial position in the turn unit.

Feature	Level	$\hat{P}$(variant = *why…for*)
Clause Type	matrix	0.72
	relative	0.36
	embedded	0.17
Subject Type	second person	0.75
	third person	0.52
	first person	0.51
	omitted	0.48
	other	0.41
Progressive Aspect	progressive	0.91
	nonprogressive	0.74
Polarity	positive	0.74
	negative	0.22
Direct Speech	direct	0.82
	regular	0.74
Turn Position	initial	0.82
	noninitial	0.74
Tense	present	0.71
	omitted	0.68
	past	0.60
	untensed	0.57

Table 4: Partial dependency information on influential qualitative features in the *why/why…for* model

Relatedly, Figure 3 shows the predicted probability of *why…for* for increasing values of the (scaled) intervening-distance variable. This demonstrates that *why…for* is favoured in clauses that have relatively little intervening material between the offset of the interrogative and *for*'s gap.

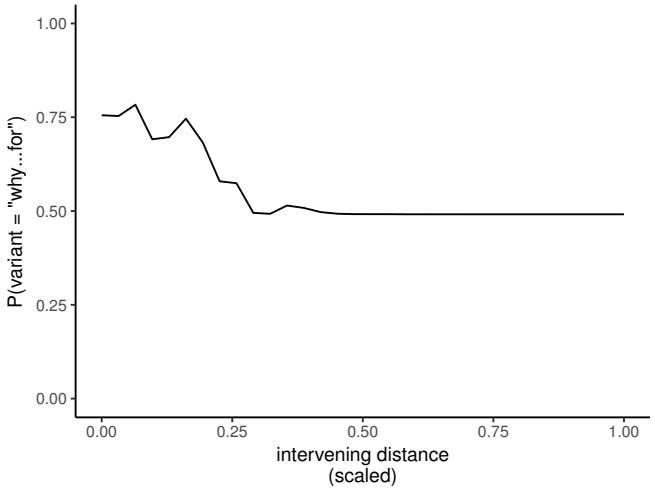


Figure 3: Partial dependency on the intervening distance feature in the *why/why…for* model

To concretize the above, consider the examples in 19 and 20. Example 19 is what one might call—in the parlance of Gries (2003) and Bernaisch et al. (2014)—a "prototypical" *why…for* construction, as it has a relatively high probability for this variant. Specifically, it is situated in a matrix clause with a second person subject, the tense marker is omitted, the IP has progressive aspect, the clause has positive polarity, and there are just two intervening words between the interrogative and *for*.

(19) can't get that side plumb . can you come and help me please. cos this. this. this plumb is like showing directions that way. not on that side but. **why you lying for**?

By contrast, 20 has a relatively low probability of *why…for*, and is non-prototypical: a larger number of words (=6) intervene between the interrogative and *for*, the clause has a first person subject, and the clause is non-progressive in its grammatical aspect. Apart from that, however, the observation is in a matrix clause with positive polarity—both favouring environments for *why…for*, hence presumably why it was realized as such.

(20) know erm fucking erm job centre . but i'm not doing it i don't see the point in doing it cos i don't wanna do it they're sitting there and talking about c . getting to know each other i'm like "fuck that" i ain't doing that . like

why do i wanna know these people for it's not as if i'm gonna chat to them outside college do you get me?

5.2.2 *Why* vs. *what…for*

Figure 4 shows that the alternation between *why/what…for* is chiefly governed by two linguistic features, progressive aspect and subject type. The plot shows a long tail of other predictors whose variable importance, on average, is greater than zero, several of which are relevant also to the *why/why…for* alternation (e.g., intervening distance, polarity, clause type, tense). However, as indicated by the values to the right of each bar, only progressive aspect ($= 1$) and subject type ($= 0.97$) are found to be relevant in over 95% of the random forest iterations, and so according to our strict criterion we should exclude them from further interpretation.

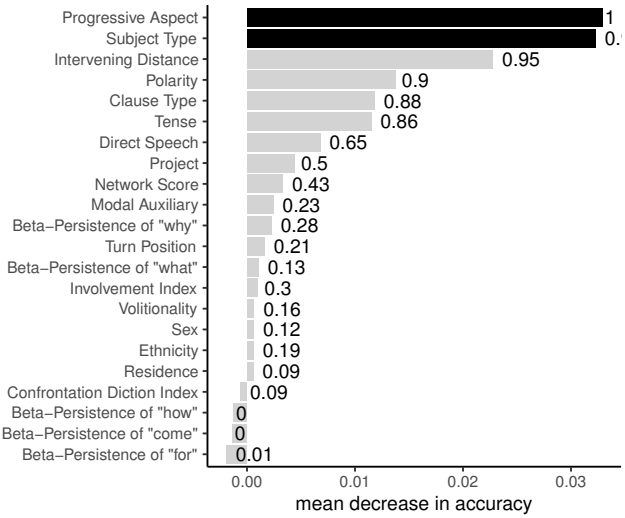


Figure 4: Variable importances for features in the *why/what…for* model. Features with bars shaded in black are deemed to be of relevance. The values to the right of each bar give the proportion of times the feature was deemed to be relevant out of 100 forest iterations.

Table 5 shows that clauses with progressive aspect and second person subjects increase the probability of a *what…for* outcome. Note that the effect of these two predictors for this part of the alternation is similar to their effect for the *why/why…for* contrast.

Feature	Level	$\hat{P}$(variant = *what…for*)
Progressive Aspect	progressive	0.80
	nonprogressive	0.02
Subject Type	second person	0.46
	other	0.19
	third person	0.09
	omitted	0.04
	first person	0.02

Table 5: Partial dependency information on qualitative features in the *why/what…for* model

5.2.3 *Why* vs. *how come*

Figure 5 and Table 6 present variable importances and partial effect directions for the *why/how come* model, respectively. These show that only clause type and polarity matter to this part of the alternation, with matrix clauses with negative polarity favouring *how come*.

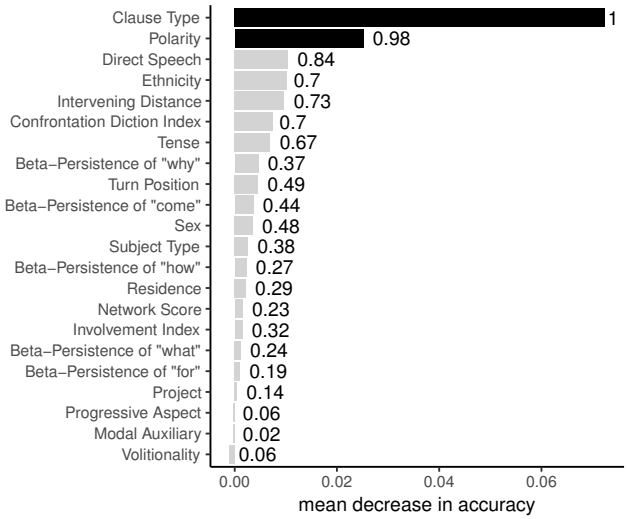


Figure 5: Variable importances for features in the *why/how come* model. Features with bars shaded in black are deemed to be of relevance. The values to the right of each bar give the proportion of times the feature was deemed to be relevant out of 100 forest iterations.

Feature	Level	$\hat{P}$(variant = *how come*)
Clause Type	matrix	0.50
	relative	0.26
	embedded	0.01
Polarity	negative	0.99
	positive	0.49

Table 6: Partial dependency information on qualitative features in the *why/how come* model

5.2.4 *Why…for* vs. *how come*

Three features strongly distinguish *why…for* from *how come*—namely, polarity, progressive aspect, and tense (Figure 6). Specifically, *why…for* is favoured over *how come* when the clause has positive polarity, progressive aspect, and an omitted tense marker (auxiliary) (Table 7).

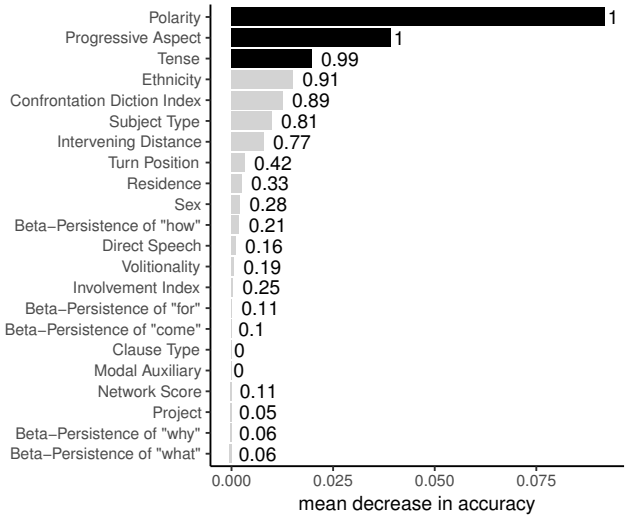


Figure 6: Variable importances for features in the *why…for/how come* model. Features with bars shaded in black are deemed to be of relevance. The values to the right of each bar give the proportion of times the feature was deemed to be relevant out of 100 forest iterations.

Feature	Level	$\hat{P}$(variant = *why…for*)
Polarity	positive	0.48
	negative	0.01
Progressive Aspect	progressive	0.93
	nonprogressive	0.49
Tense	omitted	0.79
	past	0.64
	present	0.48

Table 7: Partial dependency information on qualitative features in the *why…for/how come* model

5.2.5 *What…for* vs. *how come*

Finally, Figure 7 and Table 8/Figure 8 present variable importances and partial effect information for the *what…for/how come* contrast, respectively. For this part of the alternation, polarity, progressive aspect, tense, clause type, and intervening distance appear to be relevant. In contrast to *how come*, *what…for* is favoured in embedded clauses, with positive polarity, progressive aspect, and omitted tense marking.

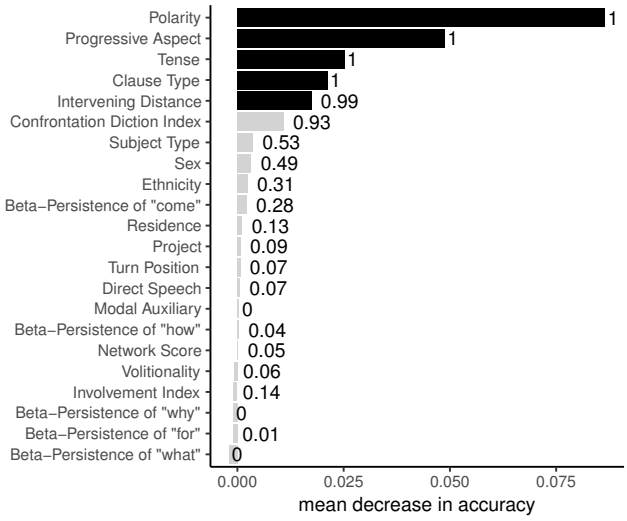


Figure 7: Variable importances for features in the *what…for/how come* model. Features with bars shaded in black are deemed to be of relevance. The values to the right of each bar give the proportion of times the feature was deemed to be relevant out of 100 forest iterations.

Feature	Level	$\hat{P}$(variant = *what…for*)
Polarity	positive	0.14
	negative	0.00
Progressive Aspect	progressive	0.82
	nonprogressive	0.15
Tense	omitted	0.76
	past	0.53
	present	0.15
Clause Type	embedded	0.54
	matrix	0.14

Table 8: Partial dependency information on qualitative features in the *what…for/how come* model

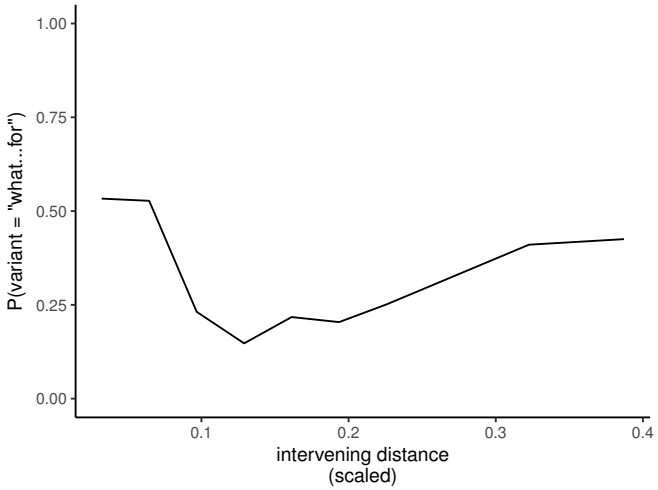


Figure 8: Partial dependency on the intervening distance feature in the *what…for/how come* model

6 Discussion

The exploratory machine learning analyses presented above have shown that it is chiefly linguistic factors and determinants of the conversational context that underpin the alternation. Table 9 summarizes the important ones. What explanations could there be for these findings? We will focus exclusively on the contrast between *why* and *why…for* for which more data were available, and focus on two explanations— syntactic and psycholinguistic.

	why/ why…for	*why/ what…for*	*why/ how come*	*why…for/ how come*	*what…for/ how come*	#
Progressive Aspect	✓	✓	—	✓	✓	4
Polarity	✓	—	✓	✓	✓	4
Clause Type	✓	—	✓	—	✓	3
Tense	✓	—	—	✓	✓	3
Subject Type	✓	✓	—	—	—	2
Intervening Distance	✓	—	—	—	✓	2
Direct Speech	✓	—	—	—	—	1
Turn Position	✓	—	—	—	—	1

Table 9: Summary of Influential Features

6.1 Syntactic explanation

We have seen that sensitivity to polarity cannot be explained by sociolinguistic factors, but perhaps can be understood from a psycholinguistic perspective. In this section we argue instead that sensitivity to polarity can be understood through the lens of universal grammar, in that different syntactic structures for different interrogatives (which are attested cross-linguistically) can explain the categorical lack of negation in *why…for* clauses.

Earlier we discussed the differences in the syntactic distribution of *why*, *what…for*, and *how come* in English, and noted that it has been argued that reason and purpose interrogatives should be separated out in some cases, although they fall under the same causal interrogative umbrella (Zwicky and Zwicky 1971, Huddleston and Pullum 2002). We review some further cross-linguistic evidence for the syntactic separation of the two, and argue that *why* and *why…for* in MLE mark different types of causal interrogative: *why* is ambiguous between a reason and purpose question, but *why…for* is an unambiguous reason form. Following Stepanov and Tsai (2008), we propose that these differences can be understood in terms of the merge position of the different interrogatives. Predictions are made about the behaviour of the different interrogatives in different environments, and acceptability judgments show that those predictions are borne out. Therefore, we propose that the difference in syntactic structure between *why* and *why…for* is what produces the categorical absence of negative polarity in *why…for* clauses.

6.1.1 Reason and purpose across languages

It has been observed that a number of languages morphologically distinguish two different forms of 'why': one which questions the reason for some event, and the other which questions the purpose of an agent in carrying out the event. For example, in Russian, purpose-*why* questions are introduced by *začem* (21), whereas

reason-*why* questions are introduced by *počemu* (22; Stepanov and Tsai 2008).

(21) a. *Začem Ivan sjuda pišel?*
why^{P} Ivan here came
'For what purpose did Ivan come here?'

b. *Čtoby kupit' pivo*
in.order.to buy beer
'To buy beer'

(22) a. *Počemu Ivan sjuda prišel?*
why^{R} Ivan here come
'Why did Ivan come here?'

b. *Potomu čto emu bylo skučno*
because him was boring
'Because he was bored'

The purpose-*why* question and the reason-*why* question forms are felicitously answered by giving a purpose, and a reason, respectively. Stepanov and Tsai (following Tsai 2008) also point out that Mandarin Chinese shows a similar distinction, where *weishenme* can have a reason or a purpose reading, but *wei-le shenme* (with a fossilized aspect marker *le* between *wei* and *shenme*) can only have a purpose reading. Interestingly, the two different *why*s in both languages can also be distinguished by their syntactic behaviour, particularly with respect to their interactions with various scope taking elements in the clause, such as modals and negation. In Russian, purpose-*why začem* is infelicitous with negation (23a), but reason-*why počemu* is fine in the same environment (23b).

(23) a. **Začem vy ne skazali mne ob etom?*
why^{P} you not said me.DAT about this

b. *Počemu vy ne skazali mne ob etom?*
why^{R} you not said me.DAT about this
'Why didn't you tell me about this?'

In Mandarin, this contrast does not hold, and either form can appear with negation (24). However, there is a distributional difference with respect to word order: reason-*why* has to appear to the left of modals (25a,b), but purpose-*why* has to appear to their right (25c,d).

(24) a. *Akiu weishenme bu hui likai?*
Akiu why not will leave
'Why will Akiu not leave?'

b. *Akiu bu hui wei-le shenme likai?*
Akiu not will for.what leave
'What is the purpose x such that Akiu will not leave for x?'

(25) a. *Akiu weishenme hui zou?*
Akiu why will leave

b. **Akiu hui weishenme zou?*
Akiu will why leave
'Why would Akiu leave?'

c. *Akiu hui wei(-le) shenme cizhi?*
Akiu will for(-le) what resign
'For what purpose will Akiu resign?'

d.?? *Akiu wei(-le) shenme hui cizhi?*
Akiu for(-le) what will resign

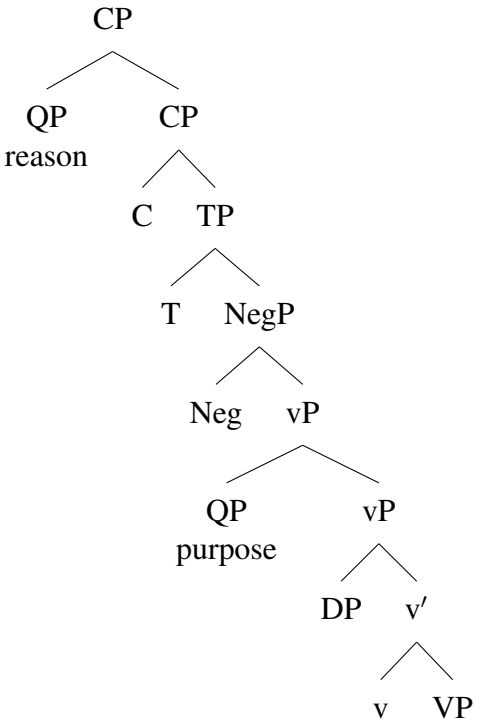
Mandarin purpose-*why* (*wei-le shenme*) furthermore exhibits an agentivity restriction, which means that it cannot appear with passives or unaccusatives (among other types of non-agentive environments, which we leave aside here).

(26) a. *na-ge xuesheng weishenme/*wei-le shenme bei pian-le*
that-CL student why/for.what BEI cheat-PRF
'Why/*for what purpose was that student cheated?'

b. *na-ben shu weishenme/*wei-le shenme chu-xian le*
that-CL book why/for.what show-up INC
'Why/*for what purpose did that book show up?'

These facts, among others related to multiple wh-questions, lead Stepanov and Tsai to conclude that there are two separate merge sites for the two different forms of 'why'. Reason-*why* must scope over at least TP and presumably merges in the CP domain, whereas purpose-*why* appears lower in the vP domain, presumably adjoined to vP. The resulting structure of the clause is something like that presented in (27).

(27)



Purpose-*why* being merged low accounts for its position relative to modals in Chinese (a robustly wh-in-situ language), and also captures the negation facts: negation is blocked with purpose questions in Russian because the purpose wh-element has to move to a higher position in the clause over Neg, which acts an intervenor (Rizzi 1990, Beck 2006, *inter alia*), but in Mandarin this effect isn't seen because licensing of wh-in-situ is available through unselective binding by a Q operator in the CP domain (Reinhart, 1998).

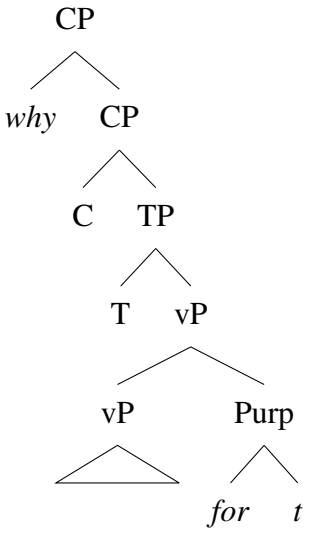
Stepanov and Tsai finish their discussion of the two *why*s by arguing that there is no such contrast between *why*s in English, and that fundamentally there is no distinct purpose-*why* in the language. We suggest in the next section that *why…for* looks like a prime candidate for an unambiguous purpose-*why* in English.

6.2 *why…for* and negation

As the discussion of the corpus data above has shown, negation is unattested in the data set in *why…for* clauses. *Why…for*'s resistance to negation could be straightforwardly explained if one were to adopt the analysis of Stepanov and Tsai (2008) discussed above, and assume that *why…for* represents a dedicated purpose-*why*

form, whereas bare *why* is ambiguous between the reason and purpose. This means that we would have the following structure for a purpose question in English:

(28)



Since English *wh*-questions are required to move to spec CP to be interpreted, when negation is present, *why* cannot move over Neg to spec CP, and the structure is ill-formed.[16]

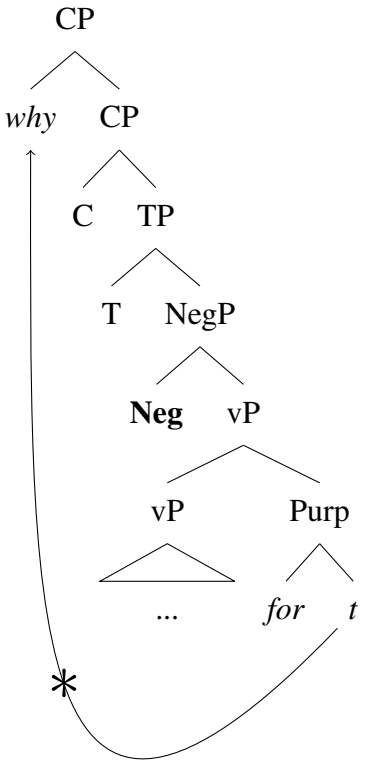
[16]Unselective binding is unavailable for English where the *wh*-element does not contain an appropriate nominal part, which provides a variable for binding. Stepanov and Tsai (2008) suggest that this is the reason for the following contrast:

(i) a. *Who left why?

b. ✓ Who left for which/what reason?

(a) is bad because *why* cannot be licensed through unselective binding, but *which/what reason*, since it contains an appropriate nominal variable, can, and so (b) is grammatical.

(29)



If *why…for* and *what…for* are unambiguously purpose forms, this means that we also expect them to behave similarly to other purpose-*why* forms crosslinguistically, with respect to agentivity, or at least, with respect to their interaction with robustly nonagentive unaccusatives and passives. As noted in section 1.1.2, Zwicky and Zwicky (1971) point out that *what…for* is unacceptable in strongly nonagentive contexts, the example given below being an unaccusative context.

(30) *What did the cake rot for?

This seems to hold for *why…for* too: the data set gave us no examples of *why…for* with robustly nonagentive unaccusative verbs or in passive constructions, and informal grammaticality judgement tasks with speakers reveal that sentences with

why…for and such unaccusative predicates are judged as ungrammatical.[17]

(31) a. Why did he have a heart attack (#for)?

b. Why did he die (#for)?

While *what…for* and *why…for* are unambiguously purpose questions, bare *why* is ambiguous between purpose and reason. On this account, bare *why* can therefore be introduced in either position in the clause (see example 27). This leads to some predictions. First, it should be the case that bare *why* is infelicitous as a purpose question when sentential negation is present. This is difficult to tell from the question itself, but Chapman and Kučerová (2016) suggest that the form of answer that is appropriate to each type of question reveals that this is indeed true. (32) is an example adapted from their work, and suggests that purpose-*why* is not possible with negation.

(32) **Context**: To put a pool in his back garden, John needs to have a wall in place to act as a support structure.
Question: Why didn't John tear down that ugly wall?

a. ✓ Because he needs it there if he wants to build a pool.

b.?? To (be able to) put in a pool later.

This contrast is expected if purpose-*why* in the form of bare *why* is also blocked from moving to spec CP by intervening Neg.

Second, we expect that bare *why* in matrix spec CP could not have originated in an embedded clause if negation is present in the matrix clause, since it would have been blocked regardless of whether it originates in the reason or purpose position. Note that bare *why* normally can be interpreted as having originated in the embedded clause where negation is not present:

(33) *Why did John say that Sally resigned?*

a. Low construal reason: ✓ *Because she was being treated unfairly.*

b. Low construal purpose: ✓ *In order to get a job.*

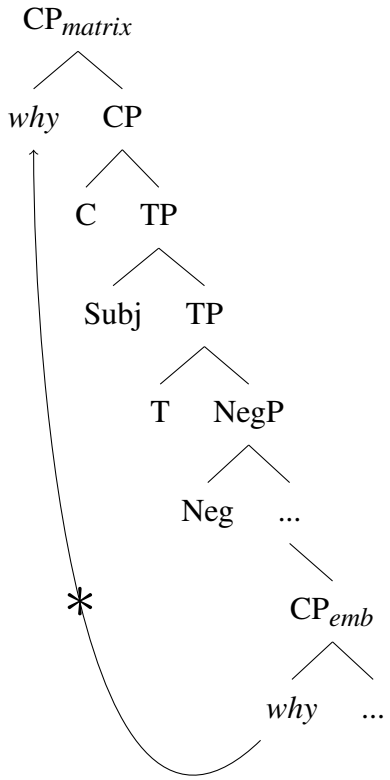
With negation in the matrix clause, both low readings are blocked, and the only available interpretations become those which involve a matrix origin of *why*.

[17] At least, if they are grammatical, they can only be interpreted with the unusual meaning that the person in question purposefully had a heart attack or died, which is why we have marked the examples with # rather than *.

(34) *Why didn't John say Sally resigned?*

This is, again, expected if negation acts as an intervenor, and blocks any movement of *why* from the embedded CP to the matrix CP, as shown in (35).

(35)



In summary then, bare *why* and *why…for* have the following properties:

(36) a. bare *why*: can be generated high or low, and thus is ambiguous between reason and purpose readings; cannot move over matrix negation from embedded CP.

b. *why…for*: always generated in low purpose position; cannot move to spec CP over negation, even in matrix clause, so is always unavailable when negation is present.

6.2.1 Potential counterexamples from the corpus

There are two examples from the corpus that could potentially be counterexamples to the claim that *why…for* is a dedicated purpose-*why*. The two examples are given in (37), and both are examples where the speaker is reporting a conversation that they had, and reporting their interlocutors response to their questions.

(37) a. "Why do you keep changing the names for?" . "because I don't know"
(Kate; Donna_Kate_1; LIC)

b. "Why you ticketing my car for?" he'll say "you parked on a double yellow like so what do you expect?"
(Hadiya; Hadiya_Isabella_Bisa_2; LIC)

In both cases, *why…for* is used, but then the answer (particularly clearly in the second example) is a reason, not a purpose.

There are two ways to confront these examples. The first is the approach taken by Stepanov and Tsai (2008) to similar counter-examples to their claims about the distinction between forms of 'why' in Russian. They suggest that any purpose can in fact be restated in terms of a reason, and be truth-conditionally equivalent, and so it is never infelicitous to answer a purpose question with a reason answer. The problem with the examples that we have here is that we are attempting to probe the nature of the question by means of the answer, which gets us into murky territory.

Take example (37b). One could imagine that the appropriate interpretation of the question is something along the lines of "what is your intention in giving me a ticket?". In that case, the response "what do you expect", is side stepping the question, but is still providing a pragmatically relevant response. It doesn't matter what the purpose or intention is; you know that if you park on a double yellow line, you will have to receive a ticket.

A second way to approach these examples is to suggest that *why…for* is also in fact ambiguous between a reason and purpose reading, and that there is variation with respect to the spell-out of the trace of the reason *why* in its high position. This approach immediately faces a problem, however. If there really was ambiguity, then we would expect to see *why…for* with negation, and we simply don't, and it is judged as infelicitous by speakers when negation is present. This suggests that the first approach is likely the right one: the answer to a purpose question can be a reason, so long as it is pragmatically relevant.

6.3 Psycholinguistic explanation

As seen above, *why…for* is relatively more frequent in the following contexts: (i) progressive aspect, (ii) positive polarity, (iii) matrix clauses, (iv) present tense (and when there's an omitted tense marking auxiliary), (v) second person subjects, (vi) and short clauses, (vii) direct speech contexts, and (viii) initial turn positions. We will suggest that the favouring contexts for *why…for* are cognitively easy environments while those disfavouring *why…for* are environments that place greater burdens on the production processor.

It is relatively easy to argue, and perhaps obvious, that short clauses should be easier to produce than longer ones, as the speaker has to access and combine fewer lexical items and build less structure. One important diagnostic for production difficulty comes from the presence of disfluencies, which the psycholinguistic literature demonstrates are used when a speaker is encountering production difficulties. For example, in the *why* portion of the dataset,[18] the presence of a disfluency up to three words prior to the interrogative clause onset is predictive of the log length of the interrogative clause according to a standard linear regression model ($\beta = 0.14809$, $s.e. = 0.03919$, $t = 3.779$, $p < 0.001$). This provides some support for the idea that longer clauses involve more processing effort than shorter ones.

It is much less clear why progressives, present tense forms, and *you*-subjects should be assumed to be cognitively easy contexts, but several pieces of evidence suggest that they might be — semantics, frequency, and disfluencies. In terms of semantics, progressives are conceptually simpler than non-progressives. Progressives focus only on the durative aspect of an event, not both its duration (however small) and culmination. Hence, they are simpler in their event structure. Present tense forms and second person subjects are semantically in the *hic et nunc*, compared with past tense forms and third person and NP subjects. That is, forms that encode reference to the *hic et nunc* presumably require less cognitive resources to process than those that are temporally or spatially displaced. In terms of frequency, in an independent corpus (BNC, demographic dialogue sample), progressive aspect and second person subjects are relatively more frequent in WHY-interrogatives than they are elsewhere. Specifically, progressive aspect structures occur in approximately 14% of WHY-interrogative clauses, but occur in only 0.39% of other clausal environments. Second person subjects occur as the subject choice 52.64% of the time in WHY-clauses, compared to 22.18% elsewhere. In contrast, third person subjects occur 32.29% of the time in WHY-clauses, and 37.94%. Present tense forms occur in 75.39% of why clauses, 72.31% in other environments. If these lines of argumentation are on the right track, we should find that disfluencies are relatively

[18] We restrict the data analysis to the *why* variants here to avoid the alternation itself confounding the analysis.

more likely before clauses that do **not** have second person subjects, progressive verb forms, and present tense. And this is exactly what we find: $\beta = -0.54861,\ s.e. = 0.20109,\ z = -2.728,\ p < 0.01$; $\beta = -0.7959,\ s.e. = 0.2807,\ z = -2.835,\ p < 0.01$; $\beta = -0.44221,\ s.e. = 0.17536,\ z = -2.522,\ p < 0.05$.

Thus, these pieces of evidence taken together are suggestive of less involved production processing for progressives, positive polarity, matrix clauses, present tense, second person subjects, and short intervening distances. But just why should cognitively easy environments such as these be favourable for *why . . . for*?

We suggest this relates to cognitive tracking/working memory. Assuming that *for* is planned with *why* (i.e. the choice is made at the interrogative site, rather than just sticking *for* on at the end), the *for* element has to be maintained in working memory until its first available insertion point. Increasing the distance between the interrogative and its first available insertion point prolongs the amount of time the speaker has to hold *for* in memory. Given that the speaker is involved with other production processing tasks, tracking *for* for too long may place additional burdens on the processor, and thus over long material it eventually gets dropped to allow other production tasks to be more efficiently accomplished. As we saw in section 6.1, bare *why* is ambiguous between a purpose and reason interrogative anyway, so dropping *for* under a heavy processing load leads to no loss of information.

7 A Brief History of *why. . . for*

As we originally assumed that the multicultural environment of inner city London might have given rise to *why. . . for*, it is somewhat surprising that sociodemographic attributes play little role in determining the variability, particularly the speaker's residence (inner city London vs. Havering). Thus, given its availability in both locations and the fact that it doesn't appear to be behaving like a canonical linguistic innovation, *why. . . for* is likely to have a different, and potentially much earlier, source. In this section, we briefly delve into the history of the variant.

The most obvious precursor of modern day *why. . . for* we can find is *forwhy* (or *for why*), with the first example attested c. 1050 according to the *OED* (see 38). This form seems to have been available through Old English to Middle English in both matrix and embedded environments (Mitchell, 1985). In this period, *forwhy* occurs with the *for* and the *why* components given in that order.

(38) Þu, Iordanen, **for hwi** gengdest on bæcling?

The first example of *why. . . for* exhibiting discontinuity occurs at the end of the 17th century in a play called *The Novelty* (1697) written by Peter Anthony Motteux (1660–1718) (see 39):

(39) Nay, **why** d'you kneel to me **for**? I a'n't your God-father.

The next discontinuous example we find is from the *Old Bailey Corpus* from 1910 (see 40):

(40) I am a coal porter. At 9 a. m. I went over the street to McKen and asked what he had been doing with the old man's things. With that he started abusing me. I struck. He took his coat off, gave it to his wife, and we had a fight, which lasted five or ten minutes. I knocked him down once or twice, and we parted. I was working for my stepfather cutting up some greenstuff. He lives at the bottom of the street. There is no thoroughfare. McKen had been to Holloway, and coming back he passed the shop, and I asked if he had been to the house and shifted a mattress. I said don't touch anything in that room. He then used bad words again and hit me a violent blow on the jaw, and I fell and lay there. He struck first. When I came to, I looked round, and he was gone. With that I walked towards where he was going, and I see him. With that my nephew see me. He said, "What is the matter with your jaw." McKen was three or four yards away. I said to him, "**Why** did you want to run away so quick **for**?" Then he sparred up again. I do not know whether I hit him or not. I fell over once; then the constable came and I went round home to the shop. At 4 in the morning they arrested me in bed. I did not kick prosecutor. All the blows I struck were in fair fighting.

While the preceding examples are from British English, the *Corpus of Historical American English* (COHA) has instances that show its availability in American English. String adjacent examples of *why for* in elliptical contexts show up from 1869 (*Dotty Dimple's Flyaway*, Sophie May) (ex. 41), with the first discontinuous example appearing in 1897 (*Wolfville*, Alfred Henry Lewis) (ex. 42), which is in an embedded clausal environment. The first attested example of *why...for* occurring in a matrix clause in American English occurs in 1904 (*A Woman's Will*, Anne Warner) (ex. 43).

(41) Didn't say a word, and the prayer-man kep' a-talkin' all the time; **why for**?

(42) An' he tells' em what he thinks an' **why** he thinks it **for**.

(43) **Why** did he want to make all that trouble **for**?

Examples 39–42, and the following one (*Mike Flannery – On Duty and Off*, Ellis Parker Butler 1909), show that *why…for* typically occurs in the context of informal, non-standard language.

(44) "And would be ye makin' poor Mike Flannery pay good money for thim rascal fleas he kilt, and him with his ankles so bit up they look like the small-pox, to say nothin' of other folks which is th' same?" she cried. "'Tis ashamed ye should be, Mister Professor, bringin' fleas into America and lettin' them run loose! Ye should muzzle thim, Mister Professor, if ye would turn thim out to pasture in the boardin'-house of a poor widdy woman, and no end of trouble, and worry, and every one sayin', '**Why** did ye let th' Dago come **for**, annyhow?"'

The *Corpus of Contemporary American English* (COCA) shows that it is well attested in recent American English (1990–), and a cursory examination of the *International Corpus of English* (ICE) points its presence in other Englishes as well.

Back in the UK, *why…for* is robustly attested in 1980s/1990s British English in the BNC. However, it is only found in spoken demographic component, further pointing to the informal and colloquial nature of this variant.

Where did *why…for* come from? Given the above survey, it is quite possible that modern day *why…for* is a direct continuation of OE *forwhy*. At some point, presumably during the growth of preposition stranding in Middle English, *forwhy* alternated with *why…for*, with the latter eventually becoming fixed (see Claridge (2012) for a similar development of *what…for*).

8 Conclusion

We began by considering the possibility that the appearance of a non-standard causal interrogative *why…for*, and the apparently variable use of the form and other forms of causal interrogative in the speech of young MLE speakers might be amenable to a sociolinguistic explanation. We have shown, using state-of-the-art data mining techniques, that variation in the use of different causal interrogatives, *why*, *what…for*, *why…for*, and *how…come* likely cannot be explained by any standard external social factors. Instead, we proposed two internal linguistic explanations for the distribution and variable frequency of different forms.

First, differences in the syntactic distribution and interpretation of questions with *why* and *why…for* can be explained by different syntactic positions for the two interrogatives. There is a cross-linguistic tendency for languages to distinguish reason and purpose interrogatives, and the difference between *why* and *why…for*

seems to lie exactly here. Simply put, *why* is ambiguous but *why…for* is used solely for purpose. A standard syntactic analysis of the difference between reason and purpose questions explains why negation is categorically absent in *why…for* contexts: *why…for* is generated below negation in the clausal spine, and moving over it represents a weak island violation, leading to ungrammaticality.

Second, differences in favouring context and frequency of use of *why* over *why…for* can be explained through a simple psycholinguistic processing preference for *why*. We argued that the favouring environments for *why…for* all involve environments that are easier to process. *Why…for* involves a long distance dependency between the two elements of the construction, meaning that the speaker has to keep track of the *for* element until it reaches an appropriate insertion point. Since bare *why* is ambiguous anyway in its interpretation, dropping *for* where a sentence is difficult to process results in no loss of information even when a purpose meaning is planned, and so we expect to see *why* more often when the speaker is under a heavy processing load.

Sometimes viewing what at first glance appears to be sociolinguistic variation through the lens of universal grammar provides us with a better explanation of the facts. This is not to say that we can categorically rule out the role of any external factors on the variation we see here, as there could be undetected influences of sociolinguistic factors on use. However, in this case, what is clear is that the main source of variation in use is internal cognitive systems.

References

Beck, S. (2006), 'Intervention effects follow from focus interpretation', *Natural Language Semantics* **14**(1), 1–56.

Bernaisch, T., Gries, S. T. and Mukherjee, J. (2014), 'The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes', *English World-Wide* **35**(1), 7–31.

Biber, D. (1991), *Variation across speech and writing*, Cambridge University Press.

Bird, S., Klein, E. and Loper, E. (2009), *Natural Language Processing with Python*, O'Reilly Media.

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**, 5–32.

Bresnan, J., Cueni, A., Nikitina, T., Baayen, R. H. et al. (2007), 'Predicting the dative alternation', *Cognitive foundations of interpretation* pp. 69–94.

Bresnan, J. W. (1972), Theory of complementation in English syntax., PhD thesis, Massachusetts Institute of Technology.

Chapman, C. and Kučerová, I. (2016), Structural and semantic ambiguity of *why*-questions. Poster session presented at the Annual Meeting of the Linguistic Society of America.

Cheshire, J. (2013), 'Grammaticalisation in social context: The emergence of a new English pronoun', *Journal of Sociolinguistics* **17**(5), 608–633.

Cheshire, J., Adger, D. and Fox, S. (2013), 'Relative *who* and the actuation problem', *Lingua* **126**, 51–77.

Cheshire, J., Kerswill, P., Fox, S. and Torgersen, E. (2011), 'Contact, the feature pool and the speech community: The emergence of Multicultural London English', *Journal of Sociolinguistics* **15**, 151–196.

Cheshire, J., Nortier, J. and Adger, D. (2015), 'Emerging multiethnolects in Europe', *Queen Mary's Occasional Papers Advancing Linguistics* **33**, 1–27.

Claridge, C. (2012), The Origins of How Come and What...For, *in* I. Hegedüs and A. Fodor, eds, 'Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23-27 August 2010', Benjamins, Amsterdam, pp. 177–195.

Coveney, A. and Dekhissi, L. (2017), Conflicting rhetorical questions in multicultural communities of London and Paris, *in* H. Tyne, M. Bilger, P. Cappeau and E. Guerin, eds, 'La variation en question(s)', Bruxelles: Peter Lang.

Gabrielatos, C., Torgersen, E. N., Hoffmann, S. and Fox, S. (2010), 'A corpus-based sociolinguistic study of indefinite article forms in London English', *Journal of English Linguistics* **38**(4), 297–334.

Gries, S. T. (2003), 'Towards a corpus-based identification of prototypical instances of constructions', *Annual Review of Cognitive Linguistics* **1**, 1–27.

Hastie, T., Friedman, J. and Tibshirani, R. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York.

Hilpert, M. (2008), 'The English comparative – language structure and language use', *English Language and Linguistics* **12**, 395–417.

Huddleston, R. and Pullum, G. K. (2002), *The Cambridge Grammar of the English Language*, Cambridge University Press, Cambridge.

Hutchby, I. (2013), *Confrontation talk: Arguments, asymmetries, and power on talk radio*, Routledge.

Kerswill, P., Cheshire, J., Fox, S. and Torgersen, E. (2004–2007), 'Lnguistic Innovators: the English of adolescents in London: Full research report ESRC end of award report, res-000-23-0680'.

Kerswill, P., Cheshire, J., Fox, S. and Torgersen, E. (2007–2010), 'Multicultural London English: The emergence, acquisition and diffusion of a new variety. ESRC research project, res-062-23-0814'.

Kerswill, P., Cheshire, J., Fox, S. and Torgersen, E. (2013), English as a contact language: the role of children and adolescents, *in* 'English as a Contact Language', Studies in English Language, Cambridge University Press, pp. 258–282.

MacKenzie, L. (2013), 'Variation in English auxiliary realization: A new take on contraction', *Language Variation and Change* **25**, 17–41.

McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*, Monographs on Statistics and Applied Probability, Chapman and Hall, London.

McKinney, W. (2010), Data structures for statistical computing in Python, *in* S. van der Walt and J. Millman, eds, 'Proceedings of the 9th Python in Science Conference', pp. 51 – 56.

Mitchell, B. (1985), *Old English Syntax*, Oxford University Press.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985), *A Comprehensive Grammar of the English Language*, Longman, London.

Reinhart, T. (1998), '*Wh*-in situ in the framework of the Minimalist Program', *Natural Language Semantics* **6**, 29–56.

Rizzi, L. (1990), *Relativized Minimality*, Cambridge MA: MIT Press.

Stepanov, A. and Tsai, W. (2008), 'Cartography and licensing of *wh*-adjuncts: a cross-linguistic perspective', *Natural Language & Linguistic Theory* **26**, 589–638.

Strobl, C., Malley, J. and Tutz, G. (2009), 'An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests', *Psychological Methods* **14**(4), 323–348.

Szmrecsanyi, B. (2006), *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*, Vol. 177, Walter de Gruyter.

Tagliamonte, S. A. (2006), *Analysing Sociolinguistic Variation*, Key Topics in Sociolinguistics, Cambridge University Press, Cambridge.

Tagliamonte, S. A. and Baayen, R. H. (2012), 'Models, forests, and trees of York english: Was/were variation as a case study for statistical practice', *Language Variation and Change* **24**, 135–178.

Tsai, W. (2008), 'Left periphery and *how-why* alternations', *Journal of East Asian Linguistics* **17**, 83–115.

Vapnik, V. N. (2000), *The Nature of Statistical Learning Theory*, Statistics for Engineering and Information Science, Springer, New York.

Zwicky, A. M. and Zwicky, A. D. (1971), *'How come' and 'what for'*, ERIC.