Language and Cognition (2022), 14: 2, 249–274
doi:10.1017/langcog.2022.7

ARTICLE

Communicative efficiency and the Principle of No Synonymy: predictability effects and the variation of *want to* and *wanna*

Natalia Levshina[1]* and David Lorenz[2]

[1]Max Planck Institute for Psycholinguistics; [2]University of Rostock
*Corresponding author. Email: natalevs@gmail.com

(Received 30 July 2021; Revised 21 March 2022; Accepted 21 March 2022)

Abstract

There is ample psycholinguistic evidence that speakers behave efficiently, using shorter and less effortful constructions when the meaning is more predictable, and longer and more effortful ones when it is less predictable. However, the Principle of No Synonymy requires that all formally distinct variants should also be functionally different. The question is how much two related constructions should overlap semantically and pragmatically in order to be used for the purposes of efficient communication. The case study focuses on *want to* + Infinitive and its reduced variant with *wanna*, which have different stylistic and sociolinguistic connotations. Bayesian mixed-effects regression modelling based on the spoken part of the British National Corpus reveals a very limited effect of efficiency: predictability increases the chances of the reduced variant only in fast speech. We conclude that efficient use of more and less effortful variants is restricted when two variants are associated with different registers or styles. This paper also pursues a methodological goal regarding missing values in speech corpora. We impute missing data based on the existing values. A comparison of regression models with and without imputed values reveals similar tendencies. This means that imputation is useful for dealing with missing values in corpora.

Keywords: contraction; efficiency; predictability; Principle of No Synonymy; register and style; missing data imputation; Bayesian regression

1. Aims of this study

It has been frequently argued that language users tend to reduce the cost-to-benefit ratio during language use (Gibson et al., 2019; Hawkins, 2004; Jaeger & Tily, 2011; Levshina, 2018; Levshina & Moran, 2021). An important strategy that helps to use language more efficiently is to choose less costly forms to express more predictable (accessible, typical, frequent, discourse-given, etc.) information, and more costly forms to express less predictable information. One form is less costly than another if its duration is shorter and/or it requires less articulation effort, which depends on the

number of segments, amount of articulatory detail and prosodic prominence. Note that duration and articulation effort are correlated. Using less costly forms for predictable meanings is possible because language users know that they can rely on the interlocutor's ability to infer relevant information from linguistic cues and context under the assumption of cooperative efficient behaviour (Levinson, 2000; Levshina, 2018). This ability is based on the mechanisms of social cognition and theory of mind, although in many cases, the efficient choices become conventional and automatic as a result of repeated use (Bybee, 2010; Diessel, 2019).

Examples of efficient use of shorter and longer forms can be found in the lexicon (Mahowald et al., 2013; Piantadosi, Tily, & Gibson, 2011; Zipf, 1949), morphosyntax (Haspelmath, 2021; Kurumada & Jaeger, 2015; Levy & Jaeger, 2007) and phonology (Cohen Priva, 2008; Hall et al., 2018; Jaeger & Buz, 2017; Seyfarth, 2014). For example, it is possible to name one and the same person using different referential expressions, for example, *she*, *the professor*, *Professor Smith* or *Professor Caroline Smith from the English department.* Their choice depends on how accessible the referent is at the given point in discourse (Ariel, 2001). In morphosyntax, there are studies of use and omission of complementiser and relativiser *that*, which demonstrate that the predictability of the clause given the matrix verb or head noun/adjective increases the chances of *that*-omission (Jaeger, 2010; Kaatari, 2016; Wasow, Jaeger, & Orr, 2011). For instance, the omission of *that* is more likely after *hope* than after *show* because the former is more commonly followed by a complement clause than the latter, for example, *I hope (that) everything will be just fine* (Jaeger, 2010). As for subject-auxiliary contractions, such as *she's* or *they have*, their rate is also determined by diverse predictability measures, for example, predictability of the subject given the auxiliary or predictability of the next verb given the subject and auxiliary (Barth, 2019; Frank & Jaeger, 2008). Phonological reduction is also determined by predictability, measured in very different ways. For example, Fowler & Housum (1987) found effects of repetition on the duration of content words in a narration. Bell et al. (2009) report a significant effect of different types of conditional probability – given the previous context or the next item, as well as word frequency and repetition, on reduction of words. Cohen Priva (2008) shows that oral and nasal stop deletion in English is influenced by the phones' average informativity (i.e., the negative log-transformed probability of a phone given all the phones that precede it in the same word, averaged across every instance of the phone in the corpus), even when frequency and context-specific predictability are controlled for (see also Seyfarth, 2014).

Importantly, the efficient use of variants presupposes their functional equivalence. However, this requirement contradicts the Principle of Contrast (Clark, 1987), which is also known as the Principle of No Synonymy (Goldberg, 1995) in Construction Grammar. It is closely related to the principle of isomorphism, or 'one meaning, one form' (Bolinger, 1977; Haiman, 1980). According to this principle, two formally distinct forms should also differ functionally. That is, they should not be fully interchangeable in a given context. More specifically, the difference can be in register (e.g., *buy* vs. *purchase*), dialect (e.g., *lorry* vs. *truck*), connotation (e.g., *curious* vs. *nosy*), construal (e.g., *The policewoman arrested the thief* vs. *The thief was arrested by the policewoman*) and so forth (Goldberg, 2019, pp. 25–26). We will use Goldberg's formulation in this paper because includes stylistic and sociolinguistic differences, which are the main focus of our study.

The Principle of No Synonymy is based on pragmatic reasoning and enabled by statistical pre-emption in language acquisition (Clark, 1987; Goldberg, 1995, 2019).

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

Speakers and addressees adhere to the principle 'What is not said, is not' (Levinson, 2000): in the presence of a salient alternative to some expression, a language user will not over-extend the meaning of this expression to include the meaning of that alternative. This is why the interpretation of 'some chocolates' is not extended to include 'all chocolates'. According to Goldberg (2019, p. 26), the fact that speaker does not need to choose between fully equivalent forms has advantages for language production because unbiased decisions are more difficult to make.[1]

There is abundant evidence that children always try to infer a contrast between two different lexical or grammatical forms. For example, a child may first use the word 'dog' to refer to cats, sheep, horses and other animals. But as soon as they learn the word 'cat', they automatically stop over-extending the label 'dog' to cats (Clark, 1987). In language change, if one source construction has two or more formally distinct variants, or if some variant appears in addition to the already existing one (due to phonological change, borrowing, etc.), the resulting constructions should divide their semantic or pragmatic 'labour' and go their own ways.

Unfortunately, there is no clear definition of what nuances qualify as a difference in meaning, and forms or structures can differ or overlap at different levels (cf. Uhrig, 2015). Laporte, Larsson, & Goulart (2021) suggest that the principle holds less reliably at low levels of formal description.[2] Moreover, linguistic variation is usually probabilistic, as has been demonstrated in numerous studies of grammatical alternations (e.g., Bresnan et al., 2007; Gries, 2003; Szmrecsanyi & Hinrichs, 2008), which means that there should be a certain degree of freedom, even if one form can be strongly preferred to another in a given context. Still, there is broad consensus that variation between alternate forms should be motivated and can therefore be analysed in terms of determining semantic, syntactic, stylistic and other factors. A good variational model is expected to discriminate well between the variants, which means it should have a low amount of random ('residual') variation.

At the same time, there is evidence that at least some alternations can be used in a communicatively efficient way as described above (e.g., Levshina, 2018), such that the less costly form is used when the meaning is more predictable from context, and the more costly form is preferred in situations where the meaning is less predictable. The assumption behind most studies of efficient communication is that the intended meaning should stay the same. But if different forms have different meanings, is this assumption tenable? This important question has not been addressed yet, as far as we know.

Importantly, formal length is not only determined by predictability. A major factor is stylistic variation. If the variants exhibit a length asymmetry, the longer variant is more likely to be preferred in formal communication and careful speech, whereas the shorter one will be more appropriate in informal communication and casual speech (e.g., Labov, 1966). Generally speaking, contracted forms are considered more appropriate in informal language, while full forms are regarded as typical of formal texts (Finegan & Biber, 2001), so contractions like *I'll, aren't* or

[1]Gardner et al. (2021) argue against production difficulties in the presence of different morphosyntactic variants. However, they do not control for the equal probability of the variants *in specific context*s, where the Principle of No Synonymy operates.

[2]Uhrig (2015, pp. 334–335) cites the example of *in the street* and *on the street* as being truly synonymous in many contexts. We might say that these variants differ on the item level (*in* vs. *on*) but are equivalent on more abstract levels (e.g., PREP-DET-N).

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

they'd are less formal than *I will*, *are not* or *they would* (cf. Daugs, 2021; Nesselhauf, 2014; see also Biber et al., 1999, pp. 1128–1132). In Japanese, when asking someone for a favour, one says *yoroshiku onegaiitashimasu* in very formal situations, *yoroshiku onegaishimasu* in less formal situations, and simply *yoroshiku* when speaking to one's friends (personal knowledge). Similarly, *help* followed by a bare infinitive is considered to be less formal than the variant with a *to*-infinitive (e.g., Rohdenburg, 1996, p. 159; see also Biber et al., 1999, pp. 736–737).

As for sociolinguistic variation, reduced forms are more common in the speech of younger people and men (Bell et al., 2003), although women may prefer reduced forms if these forms are more prestigious (Ernestus, 2014). Some reduced forms may also indicate an orientation towards a local identity (Hollmann & Siewierska, 2011; Tagliamonte & Roeder, 2009). Reduced forms often carry less overt prestige than full forms; for example, the present progressive suffix variant /ɪŋ/ ('walking') is generally associated with higher status and prestige than the lenited form /ɪn/ ('walkin') (Campbell-Kibler, 2007; Trudgill, 1974, pp. 91–93).[3] On the other hand, reduced forms can carry covert prestige by indexing group belonging and solidarity. Since these values are often associated with masculinity, and since aberrant behaviour tends to be less accepted in women than in men, women have stronger motivation to avoid non-standard reduced language (cf. Chambers & Trudgill, 1998, pp. 83–85; Romaine, 2003, pp. 103–105). The age effect can be explained by overall more conservative linguistic behaviour of older speakers, probably, due to the strength of exemplar representations of the previous forms in their memory. Moreover, young men are more prone to using non-conventional forms to mark an identity as people who do not depend on social norms and restrictions (cf. Eckert, 2008).

But this correlation between length on the one side and formality and prestige on the other side is not always observed. Although *that*-omission has been claimed to be more widely spread in informal speech than in formal language (Huddleston & Pullum, 2002, p. 953), no clear stylistic or social effects on *that*-omission were detected when numerous other factors were also controlled for (Staum, 2005; Tagliamonte, Smith, & Lawrence, 2005). Moreover, formal length can depend on the genre and text type in very specific ways. For example, in written articles with high lexical density, which can be measured as type-token ratio, journalists may prefer the shorter genitive variant *with -'s* to the longer *of*-genitive for reasons of space economy, trying to cram as much information as possible into a press text (Szmrecsanyi & Hinrichs, 2008).

In general, very little is known about the impact of style and sociolinguistic factors, as well as other functional differences, on efficient use of variants. In this paper, we want to make a step towards understanding the paradox of efficient language use under the pressure of the tendency for distinct forms to be non-exchangeable, which is captured by the Principle of No Synonymy and the likes. We hypothesise that efficient use of variants is possible when the functional differences between them are small. The greater and more salient the semantic, pragmatic and stylistic differences, the less likely that predictability will play a role in the choice between the variants. Speaking about sociolinguistic variation, we can recall Labov's (1972) degrees of indexicality, which includes indicators (non-salient sociolinguistic variants), markers (variants salient

[3]There are some counterexamples, as well. For instance, negative concord in English is longer than standard negation, and is a non-standard form.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

inside a social group) and stereotypes (variants salient inside and outside of a social group, which language users are often aware of and which often have a negative value). Using this classification, we may expect the least salient indices to be the most available for efficiency considerations, and the most salient stereotypes the least available (see Hollmann & Siewierska, 2011, pp. 47–48, for a similar proposal regarding reduction due to high token frequency). Moreover, we can expect that contractions are less salient in informal speech than in formal texts, where they are perceived as inappropriate. This means that contractions are probably more likely to be recruited for efficiency purposes in informal speech, and less likely in formal language.

In this study, we focus on the variation between *want to* and *wanna* followed by an infinitive. This alternation is illustrated with an example from an old popular song 'Girls just want to have fun'.[4] Interestingly, the song's official title includes 'want to', but the pronunciation in the recording is invariantly 'wanna'. For the sake of brevity, the alternation will be designated in this paper as WANT. The main research question is whether the variation is explained only by stylistic and sociolinguistic factors, or whether predictability also plays a role. Previous research has suggested that frequency-based measures correlate with the use of the variants (Flach, 2020; Levshina, 2018), while Krug (2000) and Lorenz (2013) show that the variation is constrained by numerous stylistic and social factors. However, these accounts have not been tested simultaneously before. We approach this question using data from the spoken component of the British National Corpus (BNC) and Bayesian mixed-effects logistic regression models. Admittedly, the BNC does not provide a full picture of sociolinguistic and stylistic variation, especially when it comes to identity construction, but its relatively large size allows us to compute the probabilistic measures, which are required for testing the predictability effects.

Surprisingly, there has as yet been no study that provides a comprehensive, multivariate usage model of the WANT alternation in British English. Krug (2000) presents a thorough investigation based on the BNC, but with a more exploratory approach to the data. Others have focused on specific aspects and/or American English (see Section 2). A second goal of this study is to fill in this gap. The previous studies show us what to look out for: effects related to speech and articulation, effects of register and speech situation and factors of syntactic co-text.

We also pursue a methodological goal. In many situations, researchers are confronted with missing values in the data. For example, the BNC contains files with full demographic information about the speakers, and some others where this information is not available. In particular, for some instances of WANT in our dataset, we do not have information about the speaker's gender and age. One faces a dilemma: either to discard the incomplete data points, or to exclude the variables with missing values. Both options are suboptimal. In this paper, we explore a solution, which is known as data imputation, where the algorithm computes the missing values based on the existing ones. We compare two Bayesian regression models. One is based on the smaller dataset with complete observations only. The other one uses the full dataset with imputed gender, age group and speech rate (see more information below).

The remaining part of the article is as follows. In Section 2, we summarise previous accounts of the alternation. Section 3 describes the data and variables, and provides a

[4]1983. Written by Robert Hazard, performed by Cyndi Lauper. From the Album *She's so unusual*. Portrait Records.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

description of the imputation method and Bayesian regression modelling. Section 4 reports and compares the regression models with and without data imputation. Finally, Section 5 offers a discussion of the results. We used R, version 4.0.2 (R Core Team, 2020) for data analysis.

2. Previous research

The contraction of *want to* to *wanna* sits at the crossroads of phonetic reduction, morpho-syntactic restructuring and alternation of modal items. It has been viewed from all these angles, and in various theoretical frameworks. Most notably in generative grammar and trace theory, the syntactic conditions (i.e., 'rules') for the occurrence of *wanna* garnered much attention (cf. Falk, 2007; Lakoff, 1970; Postal & Pullum, 1982; Pullum, 1997). However, our focus will be on variation in variable contexts, namely constructions of the type *want to/wanna* V_{INF}, where the implicit subject of V_{INF} is coreferential with that of *want* or *wanna*. This perspective has been taken by usage-based studies such as Krug (2000) on British English and Lorenz (2013) on American English.

As a contracted item, *wanna* has its source in articulatory reduction, leading (perhaps gradually) to a realisation /wɒnə/ for the string *want to*. It can be seen as a case of extreme reduction of a specific sequence ('special reduction', Bybee, File-Muriel, & Napoleão de Souza, 2016), due to the high frequency and internal bondedness of *want to* (cf. Krug, 2000, p. 139). Morphosyntactically, the restructuring is from *want* + *to*-infinitive to *wanna* + bare infinitive, likewise following from the fusion of the invariant sequence *want to*, while the infinitive verb form remains an open slot in the construction (cf. Bybee, 2010, p. 43; Hudson, 2006, p. 609). With a bare infinitive complement (and its lack of inflected forms), *wanna* is structurally more 'modal-like', as also suggested by some of its usage tendencies (e.g., its dispreference after modals, cf. Krug, 2001).

Synchronically, the use of *wanna* is rather a variant choice than a case of online reduction or contraction (cf. Broadbent & Sifaki, 2013; Sag & Fodor, 1994), though there is still some gradience of variants in speech, such as [wɒnə], [wɒnɾə], [wɒntə] (cf. Bolinger, 1981; Ellis, 2002, p. 331; Lorenz, 2013, pp. 101–102). The choice is that of a modal item expressing volition, which can extend into intention or obligation (cf. Krug, 2000, pp. 147–149). Thus, in Krug's (2000) analysis, *want to* is an 'emerging modal' whose fusion into *wanna* is part of its grammaticalisation (cf. also Okazaki, 2002), and starts forming a schema with other modal items of similar form such as *gonna*, *gotta*. Lorenz (2013) views these contracted forms as undergoing a process of gradual 'emancipation' by which they become conceptually independent from the respective full forms. This means that they gradually lose the traits of reduced realisation variants and can take on stable functional and communicative properties that differentiate them from the full forms. In other words, they behave according to the pragmatic principles behind the Principle of No Synonymy.

Empirical findings on the use of *wanna* attest to its status as emerging and emancipating. In data from the 'spoken' section of the BNC, Krug (2000), p. 175) observes a strong frequency increase of *wanna* relative to *want to* in apparent time, from below 20% in the age group 60+ to just over 50% in the youngest speakers. In spoken American English (Santa Barbara Corpus), the distribution over age groups stabilises at around 75%–80% for the cohorts aged 49 and younger (Lorenz, 2013,

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

p. 44). Changes in the factors of variation suggest that *wanna* is becoming a fully independent item, yet less emancipated than *gonna* or *gotta*, as some aspects of ease of articulation persist (e.g., the contraction being favoured in higher speech rates and disfavoured at phrase ends; Lorenz, 2013, pp. 104–105). Moreover, children overwhelmingly use *wanna* in variable contexts and even overuse it as a transitive verb (as in, **Who do you wanna play with you*), suggesting that children might acquire *wanna* and *want* as separate items and subsequently learn their distributional differences (Getz, 2019).

Diachronically, it seems that the usage patterns of *wanna* vs. *want to* gradually converge with those of *gonna* and *gotta*, in particular on the level of socio-pragmatics and register (Lorenz, 2020). The most prominent and consistent property of the contractions then is to mark informality and colloquialness. In Boas' (2004) constructional formalisation, 'colloquial style' is what specifies the meaning of *wanna* in addition to semantic features inherited from *want* and *to*.

To summarise, *want to* and *wanna* are entrenched as distinct units with different social and stylistic properties. Therefore, they are unlikely to be used interchangeably for efficiency purposes, at least, not by all speakers and not in all contexts.

At the same time, there are indications that some predictability measures do play a role in explaining the variation of WANT. For example, Levshina (2018) argues that verbs that have high attraction to and reliance on WANT (cf. Schmid, 2000), have higher chances of being used with *wanna*. Flach (2020) has shown, as well, that measures of association with the following item (most clearly predictability of the verb given the construction, and collostructional strength) can to an extent predict the use of contractions like *wanna*. Similarly, Mair (2017) has proposed that token frequency of WANT + V, as well as priming through preceding contractions play a role in the production of *wanna*. However, these studies did not measure the role of stylistic factors and sociolinguistic variables. The present study fills this gap, combining social and stylistic factors (in particular, age, gender, text type, speech rate and stylistic prosody of individual verbs) with different predictability measures, which are described in the next section.

3. Methodology

3.1. Corpus data and variables

The data for this study come from the spoken component of the BNC. A Python script was used to extract all instances of *wanna*, which is represented by two tokens, *wan* and *na* in the corpus, and all instances of *want* followed by *to*, together with diverse contextual information, such as wordforms, lemmas and part-of-speech tags of neighbouring words, which helped us to annotate the data for 15 potential predictor variables. The sentences in which there was no infinitive were disregarded. We also checked manually those sentences where the verb occurs only once after *wanna/want to* and excluded erroneous hits. For example, in the sentence *I wanna packet Walker crisps* the word *packet* is erroneously annotated as an infinitive. Another example is *Do you wanna big'un?*, where *big'un* is also analysed as an infinitive in the corpus. Examples like those were excluded. The dataset included 9123 observations, after removing irrelevant and problematic hits.

The variables include structural variables, sociolinguistic variables, variables related to register and text type, and variables reflecting different types of predictability. They

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

are described below, and also summarised in Table 1. The dataset is provided in the online repository (see Data Availability Statement).

3.1.1. Structural variables

The first variable represents the response variable, *wanna* (n = 2,114) or *want to* (n = 7,009). It is called *expression* in the dataset. Second, we coded the infinitive that serves as a non-finite complement of *wanna* or *want to*. We included a variable which reflected if there is a negative particle before *wanna* and *want_to*. It is called *neg_part* in the dataset. The values are 'Yes' (n = 2,034) and 'No' (n = 7,089). Another variable describes if there is a question mark at the end of the sentence. It is labelled as *question*. Its values are 'Yes' (n = 1,784) and 'No' (n = 7,339). In addition, we coded the grammatical subject of *wanna* and *want to* (variable *subject*). The values were grouped into several categories:

- *I* (including *me*, *mine* in children's speech; n = 2,958).
- *You* (including *yous* and *ya*; n = 3,336).
- *We* (including *us*; n = 845).
- *He_she* (*he* and *she*; n = 309).
- *They* (including *them*; n = 745).
- PRON (other pronouns, e.g., *everybody*, *many*, *who*; n = 243).
- Other (common and proper nouns, numerals and other nominalisations; n = 395).
- Omitted (absent in the clause; n = 285).
- Unclear (when we could not determine the subject due to insufficient context; n = 7).

3.1.2. Sociolinguistic variables

The sociolinguistic variables include different information about the speaker. One of them is called *speakerID* and represents the ID of the speaker, as provided in the corpus. The variable *sex* describes the speaker's sex ('m' male, n = 4,381, or 'f' female, n = 3,058). The variable *ageGroup* represents the speaker's age group with the following values:

- 'Ag0': 0–14 years (n = 739).
- 'Ag1': 15–24 years (n = 618).
- 'Ag2': 25–34 years (n = 1,034).
- 'Ag3': 35–44 years (n = 1,044).
- 'Ag4': 45–59 years (n = 1,545).
- 'Ag5': 60+ years (n = 641).

In many cases, age and sex were unspecified.

3.1.3. Variables related to style, text type and register

We included five variables related to style, text type and register. Two of them were already available in the BNC metainformation. The first one was the text type (variable *textType*), which had two values: conversations ('CONVRSN', n = 4,342) and other spoken text types, for example, lessons, sermons or meetings ('OTHERSP', n = 4,781). The other variable was *settingID*, which stands for the unique ID of a

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

Table 1. Variables tested in this study

Type of variable	Label	Meaning	Values
Structural variables	*expression*	the variant (response variable)	*wanna* or *want to*
	infinitive	the verb of the infinitival complement	*be*, *go*, *say*, etc.
	neg_part	if there is a negative particle before *wanna* or *want to*	*Yes* or *No*
	question	if the sentence ends with a question mark	*Yes* or *No*
	subject	the grammatical subject of *wanna* or *want to*	*I*, *you*, *we*, *he_she*, *they*, *PRON* (other pronouns), *Other* (nouns and other lexical subjects), *Omitted*, *Unclear*
Sociolinguistic variables	*speaker*	Speaker's ID	specific IDs
	sex	sex of the speaker	*f* (female) or *m* (male)
	ageGroup	age group of the speaker	*AG0*: 0–14 years *AG1*: 15–24 years *AG2*: 25–34 years *AG3*: 35–44 years *AG4*: 45–59 years *AG5*: 60+ years
Variables related to style, register, text type	*textType*	text type	*CONVRSN* (conversations) or *OTHERSP* (other spoken text types)
	settingID	ID of the conversation (setting)	specific IDs
	SpeechRate	speech rate, measured as number of phones per second	numeric
	Dim1	coordinate of the second verb on Dimension 1 of the CA, representing formal – informal contrast	numeric
	Dim2	coordinate of the second verb on Dimension 2 of the CA, representing the contrast between informative language and language for aesthetic purposes/ entertainment	numeric
Predictability-related variables for testing efficient behaviour	*Info_Verb_given_WANT*	informativity of the second verb given *wanna/want to*	numeric
	Info_WANT_given_Verb	informativity of *wanna/want to* given the second verb	numeric
	Info_WANT_given_left	informativity of *wanna/want to* given left context	numeric

conversation between individual speakers in a specific time, place and during a certain activity.

We also added three other variables based on additional analyses. In particular, we computed speech rate (variable *SpeechRate*), measured as phones per second (phon/sec) in a recording, as taken from the time-aligned Praat TextGrids for the audio edition of the Spoken BNC, made available by the Oxford University Phonetics Laboratory (Coleman, 2019; Coleman et al., 2012). Stretches that are not annotated (muted or marked as 'unclear' in the transcript) were excluded; short pauses and silences were not counted as phones but their duration was not discounted. We used the R package *rPraat* (Bořil & Skarnitzl, 2016) to work with the TextGrid files. Note that since we measured the rate for each recording as a whole, this speech rate is across the utterances in a conversation. It is not strictly an articulation rate but provides a measure of a conversation's general pace, and hence of the time pressure on speech production.[5] Speech rate is an important factor that boosts phonetic reduction (Ernestus, 2014; Raymond, Dautricourt, & Hume, 2006).

Finally, we also evaluated which text type the individual verbs that occur as infinitives after *want to* or *wanna* are associated with. This information helps to capture stylistic prosody of the verbs, and provides a finer-grained and more local operationalisation of register and text type than captured by the other variables at the global level of a speech recording. The corresponding variables are called *Dim1* and *Dim2*. These dimensions are taken from a simple Correspondence Analysis of the associations between all verbs as lemmas and all text types in the BNC (Greenacre, 2007; Levshina, 2015, Ch. 19). A part of the space is shown in Fig. 1 (due to its large size, we cannot show the entire map). The horizontal dimension can be interpreted as a contrast between formal (left) vs. informal communication (right), while the vertical dimension can be interpreted as a contrast between informative language (bottom) and language for aesthetic purposes and entertainment (top). Examining additional dimensions did not yield any interpretable results. Most spoken text types are located in the bottom right quadrant. The values of the individual verbs are provided in the online repository. Among the verbs with the greatest positive values on the horizontal dimension and negative ones on the vertical dimension are contracted forms *gonna*, *wanna*, as well as obscene and slang terms, such as *f****, *bugger, shit, sod*, *snog* and *shag*. So, we would expect more instances of *wanna* with the verbs that have similar scores – that is, which are highly informal, and do not represent aesthetic use.

3.1.4. Variables reflecting predictability

The fourth and final group of variables are three corpus-based measures that reflect different types of predictability information. In particular, we can expect *wanna* to be

[5]While within-utterance speech rates have been shown to affect the use of *wanna* as well as articulatory reduction in similar items such as *have to* in American English (Jurafsky, Bell, Gregory and Raymond, Jurafsky et al., 2001; Lorenz, 2013, p. 100; Tizón-Couto & Lorenz, 2018), the pace of longer stretches of speech may also correlate with phonetic reduction (cf. Raymond, Dautricourt, & Hume, 2006). In our data set, the mean speech rate is 7.7 phones/sec (median $\mu_{1/2} = 7.6$, $SD = 2.24$)

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

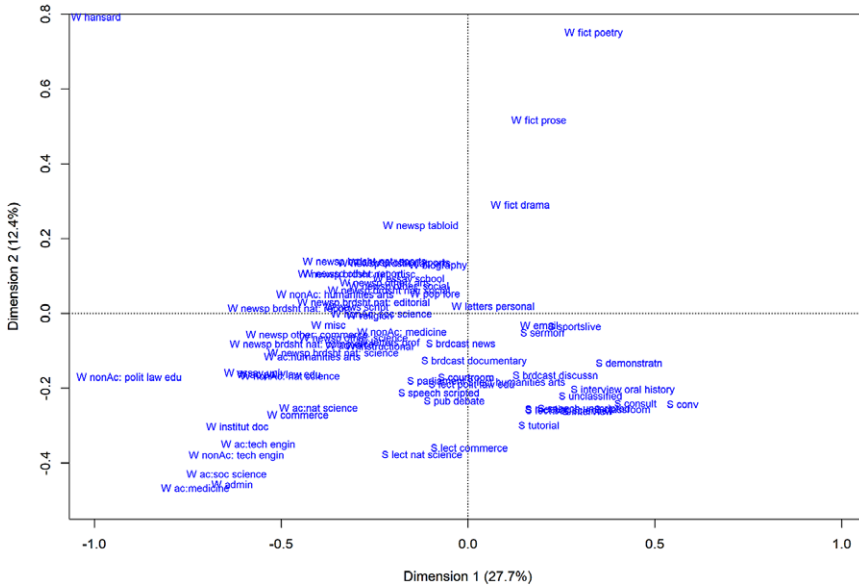

Fig. 1. A fragment of the correspondence analysis map with text categories as labels. The analysis was based on co-occurrence frequencies of verbs (not shown) and text categories. Dimension 1 can be interpreted as a contrast between formal (left) vs. informal communication (right), while the vertical dimension can be interpreted as a contrast between informative language (bottom) and language for aesthetic purposes and entertainment (top).

preferred if WANT is highly probable given the left context or the right context (the infinitive). In addition, we should test whether the probability of *wanna* is higher if the infinitive is more probable after WANT.

In accordance with previous research on communicative efficiency and formal reduction, we used informativity measures, where a negative logarithm is taken from a conditional probability. As a result, the measures represent 'unpredictability', also known as surprisal. Higher informativity means lower predictability, and the other way round. In efficient language use, high informativity is associated with longer and more effortful forms, whereas low informativity is associated with shorter and less effortful forms.

The first measure shows how unexpected an infinitive is as a complement of WANT. This variable is called *Info_Verb_given_WANT*. In order to compute this variable, we used the following formula:

$$I_{(Verb|WANT)} = -\log_2 \frac{P(Verb, WANT)}{P(WANT)} = -\log_2 \frac{F(Verb, WANT)}{F(WANT)}, \qquad (1)$$

where F (*Verb*, *WANT*) stands for the frequency of a given infinitive after *want to* or *wanna* in the data, whereas F (*WANT*) stands for the sum frequency of *want to* + Infinitive and *wanna* + Infinitive in the spoken part of the BNC.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

We also computed how unexpected *want to/wanna* is given the infinitive, using the following formula:

$$I_{(WANT|Verb)} = -\log_2 \frac{P(Verb, WANT)}{P(Verb)} = -\log_2 \frac{F(Verb, WANT)}{F(Verb)}, \quad (2)$$

where *F* (*Verb*) stands for the frequency of a given verb in the spoken part of the BNC. The reason for including this variable is the fact that backward conditional probabilities often have an effect on length (see the references in Section 1). This variable is added under the label *Info_WANT_given_Verb*.

Finally, we coded how predictable *want to* or *wanna* is given the previous word. It is called *Info_WANT_given_left* and computed as follows:

$$I_{(WANT|Word\_left)} = -\log_2 \frac{P(Word\_left, WANT)}{P(Word\_left)} = -\log_2 \frac{F(Word\_left,\ WANT)}{F(Word\_left)}, \quad (3)$$

where *F* (*Word_left, WANT*) represents the joint frequency of the word followed by *WANT*, and *F* (*Word_left*) represents the frequency of the word in the spoken component of the BNC. Note that contractions like *I'll* and *would not* were treated as words, although they are analysed as two separate tokens in the BNC. In quite a few sentences, the first word was *want* or *wanna*, followed by the infinitive, which means that there was no left context in the same sentence. For those cases, we computed the predictability of WANT to be in the beginning of the sentence, dividing the number of sentences beginning with WANT (88) by the total number of sentences in the spoken corpus (1,145,450).

We should mention here that different theories exist to explain predictability and frequency effects in formal reduction. One of them, mentioned in Section 1, assumes that the speaker/signer takes the perspective of the addressee, namely, whether the latter will be able to process the reduced form correctly on the basis of available contextual information and pragmatic principles (Jaeger, 2013; Levshina, 2018). An important role is played by information theory, which teaches how to reduce code while transmitting the message in a noisy channel (cf. Gibson et al., 2019).

In contrast, Bybee (2007, 2010) takes a predominantly speaker-centred perspective. According to her, reduction is boosted by the process of chunking of neighbouring units, based purely on surface frequency (Bybee, 2007, 2010). Each instance of use further automates and increases the fluency of the sequence, leading to fusion of the units (Bybee, 2007, p. 324). For instance, Bybee & Scheibman (1999) show that reduction of the vowel and the consonants in *do not* in spoken English is particularly frequent after the pronoun *I* and before the verbs *know* and *think*. The reason is that this contraction particularly frequently occurs in phrases *I do not know* and *I do not think*. Although Bybee (2010), p. 40) admits that the speaker controls the amount of reduction, according to the listener's needs,[6] it is the speaker-internal processes that

[6]Lorenz & Tizón-Couto (2020) likewise evoke speaker-hearer interaction and the production-perception loop in the development of contractions.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

drive the reduction. Another speaker-centred account has to do with the speaker 'buying time' to prepare a continuation with low accessibility (Ferreira & Dell, 2000).

These accounts are very difficult to disentangle. We do not know yet if *wanna* has emerged due to the fact that *want to* followed by an infinitive was highly frequent or predictable given context (e.g., the presence of an infinitive). With regard to the synchronic variation, if the chunking account is correct, we can expect *wanna* to be preferred more strongly after the subjects that are used frequently with *want to/wanna*, comparing the joint probabilities instead of conditional probabilities. This approach would also require testing the joint probabilities of WANT and verbs. However, the ranking of these probabilities is the same as the rankings of *Info_Verb_given_WANT*, which is based on the frequencies of WANT and verbs. The same holds for the 'buying time' account. If it is correct, *wanna* would be more frequent when the verb is highly probable after *want to/wanna.* At present, we cannot distinguish between these accounts because they generate the same predictions for the data that we have. We hope that this will be done in the future.

3.2. Data imputation

Unfortunately, the spoken component of the BNC does not contain all values for the sociolinguistic variables. We also did not have access to some recordings in order to compute speech rate. There was a high proportion of missing values. In particular, *sex* had 18% missing observations, *ageGroup* 38% and *SpeechRate* 47%. These are very substantial proportions. The missing values are more frequent in the 'other spoken' texts than in the conversations.

To solve this problem, we performed two analyses. The first one is based on the data without missing values. This is a small dataset with only 3,603 observations. The second method is to use all data and impute the missing values. If the missing values were few, one might get by with simply setting them to the median or reference level of the variable; but since some of our variables have many missing observations, we need a more fine-grained method (see Harrell, 2015, pp. 47–57 for further discussion). For this purpose, we used the procedure of multiple imputation implemented in the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011). This approach predicts the missing values based on the values in the other variables.

The imputation algorithm is based on so-called chained equations. A series of regression models is fitted, whereby each variable with missing data is modelled conditional upon the other variables in the data. For the binary variable (*sex*), the regression is logistic. For *ageGroup*, which is an ordered factor, it is an ordinal proportional odds model. For *SpeechRate*, it is a Bayesian linear model. The information about the setting IDs, speaker IDs and the infinitive was not considered for prediction because they had very low frequencies (median frequency of each individual level = 2), which made the computations unreliable and prohibitively slow, even on a computer cluster. The imputed speech rates were restricted to the range between 0 and 15 phones per second because this is where the 99% of the observed data lay.

In the next step, the predicted values in one of those variables are set back to missing and the procedure is repeated; this time the imputed values in the other variables are used for prediction. This procedure is repeated several times. In our analysis, we used 50 iterations. The cycles converge, fluctuating randomly around a narrow range of values.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

The imputation algorithm returned five imputed datasets (the default option), which were identical in the non-missing values, but differed in the imputed values. These differences are due to uncertainty in the Bayesian probabilistic sampling used by multiple imputation. Next, they were averaged as one dataset with the same number of observations as the initial dataset, but with the imputed values instead of the missing ones. Finally, the dataset with imputed values was used to fit the second regression model.

3.3. Bayesian GLMM

We used Bayesian mixed-effects generalised linear models (GLMM) with the logit transformation (package *brms* in R, see Bürkner, 2018). Bayesian modelling represents an attractive alternative to frequentist regression, which is also known as the maximum likelihood method. First, it allows the researcher to test the alternative hypothesis directly, rather than to test if the null hypothesis can be rejected. Second, it does not involve binary decisions based on *p*-values. Third, the generalised mixed linear models in the maximum likelihood version often have convergence issues, which can be more easily solved in the Bayesian framework. For an explanation of the principles of Bayesian inference, see Kruschke (2011), McElreath (2016) and Nicenboim & Vasishth (2016); for a comparison of frequentist and Bayesian regression in models of language variation, see Levshina (Levshina, in press).

In practical terms, the difference between frequentist and Bayesian regression is that the regression coefficients are estimated as the mean posterior distributions of the parameters given the data, rather than as point estimates based on the maximum likelihood estimation. The distributions are generated with the help of the Monte Carlo Markov Chain (MCMC) sampling. In practice, however, the coefficient estimates of maximum likelihood models and those of Bayesian models are very similar. What differs is the process of inference. The posterior distributions based on MCMC enable us to compute 95% credible intervals for an estimate, where the parameter of interest (e.g., a regression coefficient) falls with the probability of 95%. We can also compute the probability that a certain estimate has a positive or negative effect on the use of *wanna* vs. *want to* (cf. Nicenboim & Vasishth, 2016; Vasishth et al., 2013).

The technical details about the Bayesian models and their goodness-of-fit measures are provided in the Supplementary Materials. The model selection process was as follows. First, all variables listed above were tested as fixed effects, with the exception of infinitives, subjects of WANT, setting IDs and speaker IDs, which served as random intercepts. We also tested models with fine-grained text categories as random intercepts (which partly overlap with the predictor *textType*), but found that this variable did not make any substantial difference. This is why we chose the simpler models without this random effect. All pairwise interactions were tested, such that the models with them had WAIC (Watanabe-Akaike Information Criterion) less than the model without any interactions, and the difference was more than one standard error. The model based on the small sample contained an interaction between informativity of WANT given the word on the left (*Info_WANT_given_left*) and speech rate. This interaction was also found to be relevant in the model based on the large sample with imputed values. In addition, the second model contained another interaction that involved two informativity measures. They reflect the

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

predictability of WANT given the left context and the second verb: *Info_WANT_given_left* and *Info_WANT_given_Verb*.

4. Results of Bayesian modelling

The results of our modelling are shown in Table 2. It displays the coefficients of the fixed effects and their 95% credible intervals. The estimates greater than 0 mean that the condition increases the chances of *wanna*, whereas negative values indicate that the chances of *want to* increase. The right-hand column shows the posterior probability of the coefficients to be positive, computed as the proportion of positive values

Table 2. Table of coefficients of the Bayesian models

Parameter	Mean posterior	Lower boundary of 95% CI	Upper boundary of 95% CI	Probability > 0
Intercept	*−6.14*	*−8.17*	*−4.25*	*0%*
	−5.87	−7.42	−4.37	0%
textType = CONVRSN (conversations) vs. OTHERSP (other texts)	*3.47*	*2.73*	*4.25*	*100%*
	3.73	3.30	4.19	100%
neg_part = Yes (vs. No)	*−0.71*	*−1.15*	*−0.26*	*0%*
	−0.31	−0.61	−0.01	2.2%
question = Yes (vs. No)	*0.13*	*−0.16*	*0.43*	*81%*
	0.10	−0.12	0.32	80.7%
sex = m (male) vs. f (female)	*0.90*	*0.43*	*1.38*	*100%*
	1.04	0.75	1.34	100%
ageGroup = 1 (vs. 0)	*0.24*	*−0.18*	*0.68*	*85.6%*
	0.31	0.06	0.55	99.2%
ageGroup = 2 (vs. average of groups 0 and 1)	*0.07*	*−0.15*	*0.30*	*73.3%*
	−0.14	−0.27	−0.02	1.1%
ageGroup = 3 (vs. average of groups 0, 1 and 2)	*−0.30*	*−0.46*	*−0.14*	*0%*
	−0.37	−0.46	−0.28	0%
ageGroup = 4 (vs. average of groups 0–3)	*−0.21*	*−0.33*	*−0.09*	*0%*
	−0.19	−0.25	−0.13	0%
ageGroup = 5 (vs. average of groups 0–4)	*−0.21*	*−0.33*	*−0.10*	*0%*
	−0.27	−0.35	−0.19	0%
SpeechRate	*0.21*	*0.07*	*0.34*	*100%*
	0.19	0.11	0.27	100%
Info_Verb_given_WANT	*−0.01*	*−0.09*	*0.06*	*35.5%*
	−0.02	−0.07	0.04	27.5%
Info_WANT_given_Verb	*0.03*	*−0.10*	*0.17*	*67.3%*
	−0.08	−0.22	0.06	13.9%
Info_WANT_given_left	*0.13*	*−0.10*	*0.34*	*86%*
	−0.09	−0.32	0.14	21.3%
Dim 1 (informality vs. formality)	*0.13*	*−0.07*	*0.35*	*88.8%*
	0.16	0.02	0.32	98.5%
Dim 2 (aesthetics vs. information)	*−0.01*	*−0.21*	*0.19*	*45.2%*
	0.04	−0.10	0.18	71.3%
Interaction Info_WANT_given_left and SpeechRate	*−0.03*	*−0.06*	*0*	*1%*
	−0.03	−0.04	−0.01	0.1%
Interaction Info_WANT_given_left and Info_WANT_given_Verb	*NA*	*NA*	*NA*	*NA*
	0.04	0.01	0.07	99.8%

Note. The upper values in a cell (in italics) are the coefficients of the first model. The bottom values in a cell (no italics) are the coefficients of the second model based on the imputed data. Positive values mean higher chances of wanna in comparison with want to. The interaction term was absent in the first model (hence NA's), according to the results of model selection.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

in the 6,000 posterior samples. Each cell in the table contains two numbers. The upper number in italics is the estimate from the model based on the small dataset with complete observations; the lower number is related to the model based on the large dataset with imputed data. Overall, the estimates produced by the two models are similar with regard to the direction and strength of the effects, with the exception of the two informativity variables. The difference between the models is due to the additional interaction term in the second model. Also, the 95% credible intervals are narrower in the second model based on the large dataset. This is natural: Bayesian posteriors usually become more specific if more data are available.

The coefficients and probabilities tell us that *wanna* occurs more often in conversations than in the other texts. The effect is very strong and highly credible, as we can see from the 95% credible intervals, which do not include zero in either of the models. Also, the posterior probabilities that conversations boost the chances of *wanna* are 100%.

The presence of the negative particle decreases the chances of *wanna* with sufficient credibility, which means that it increases the chances of *want to*. Questions somewhat increase them, but this effect is weak, and the posterior probabilities are below 90%.

As for the sociolinguistic variables, there is a clear effect of the speaker's sex in both models. Male speakers use *wanna* more often than female speakers. Also, the use of *wanna* is less likely in the older groups than in the younger groups. Figure 2 shows the

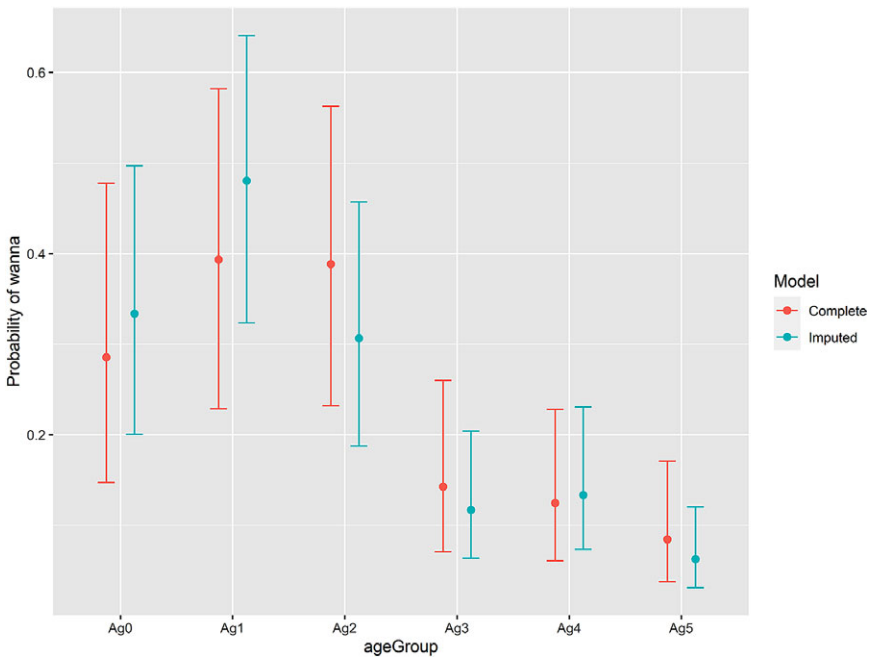

Fig. 2. Probabilities of *wanna* for different age groups in two models. This is a conditional plot showing the predicted probabilities of *wanna* depending on the age groups, computed for following values of the predictors: conversations, female speakers, no negation, not a question and mean values of the numeric predictors in the data set with imputed values. The points represent the mean posterior estimates, and the error bars stand for 95% credibility intervals based on MCMC sampling. Red: the model with complete observations only; turquoise: the model with imputed values.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

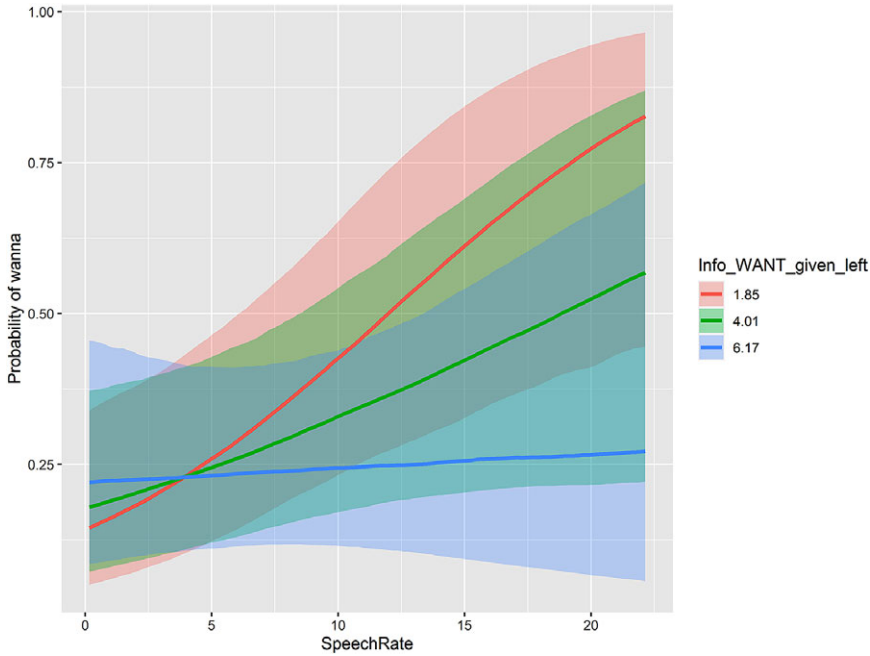

Fig. 3. Interaction between speech rate of a recording and informativity of WANT given word on the left in the model with complete observations. This is a conditional plot showing the predicted probabilities of *wanna* depending on Speech Rate and Informativity of WANT given the word on the left, computed for the following values of the predictors: conversations, female speakers, Age Group 0, no negation, not a question and mean values of the remaining numeric predictors in the dataset with complete observations only. The lines correspond to the predicted posterior mean values, and the shaded areas are 95% error bands based on MCMC sampling. The blue line corresponds to high values to the model predictions for *Info_WANT_given_left* (6.17); the green line corresponds to medium values (4.01), and the red line corresponds to low values (1.85).

probabilities of *wanna* for different age groups in each model. It is a so-called conditional plot, so the probabilities of *wanna* are computed for some selected values of the other categorical variables (i.e., conversations, female speakers, no negative particle, not a question), and mean values of the continuous variables. The plot shows that the probability of *wanna* is higher in the three younger groups than in the three older groups.

The interaction between speech rate and informativity of WANT given the word on the left is visualised in Fig. 3. It is based on the model with complete observations, but the results for the model with imputed values are very similar. By default, the three values of the informativity variable chosen for visualisation are the mean (4.01, green line) plus one standard deviation (6.17, blue line) and minus one standard deviation (1.85, red line). The plot shows that the effect of predictability becomes obvious when the speech rate is higher. In fast speech, lower informativity of WANT (represented with the red line in Fig. 3) means higher chances of *wanna*, in comparison with middle-level informativity (the green line) and high informativity (the blue line). This means that contexts where WANT is more expected contain *wanna* more frequently than contexts where WANT is less expected, which can be

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

regarded as efficient use of language. Crucially, this effect only becomes obvious in fast-paced interactions. In slow-paced communication, the chances of *wanna* are equally low for all types of contexts.

Words with low informativity (and therefore high predictability) of WANT that are particularly frequently followed by *wanna* include frequent adverbs *just* and *only*, and some contractions, such as *she'll*, *d'ya* and *gonna*. Consider examples (4–6) below. They represent very informal language, so informality and high predictability seem to be intertwined. We will return to this observation in the discussion.

(4) Cos you only wanna lose half as much you are eating twice as much as we said? (KPR, a conversation).
(5) Sinead where d'ya wanna go today? (KPE, a conversation).
(6) They gonna wanna turn out Sunday you know, or? (KSR, a conversation).

There are also differences in the effects of dimensions of Correspondence Analysis. In the first model, the effects of Dimension 1 (informality vs. formality) is in the predicted direction (informality boosts *wanna*), but the 95% credible interval includes zero, and the probability of this positive effect is less than 90%. The mean posterior of Dimension 2 is around zero. In the second model, the coordinates of verbs on Dimension 1 have a stronger and more convincing positive effect on the chances of *wanna*. This means that verbs that occur in informal texts are also more frequently used with *wanna* in speech.

It is also useful to look at the random intercepts of the grammatical subjects of WANT. Figure 4 displays their distribution (in log-odds ratios, which show their

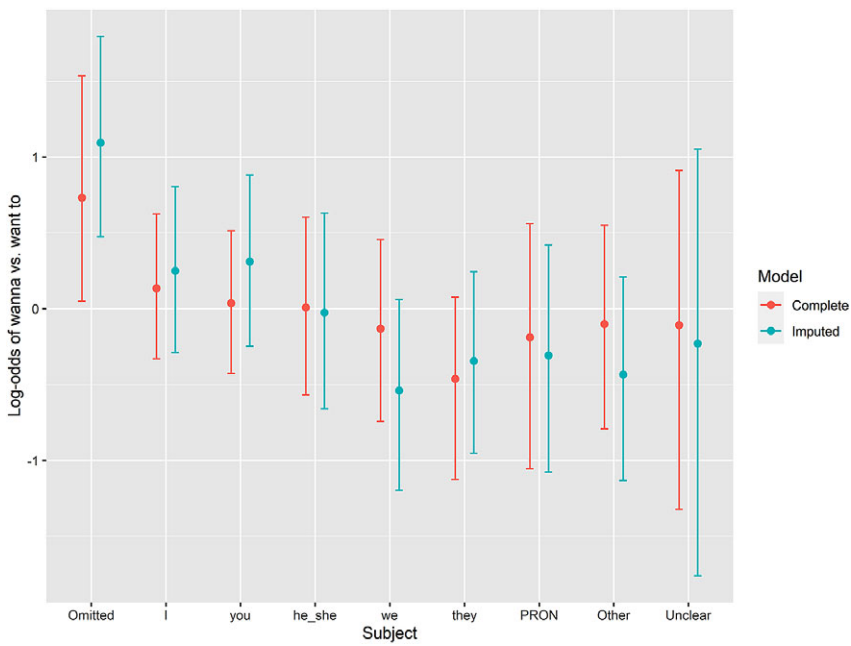

Fig 4. Random intercepts of different grammatical subjects in both models. The points represent the mean posterior estimates, and the error bars stand for 95% credibility intervals based on MCMC sampling. Red: the model with complete observations only; turquoise: the model with imputed values.

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

adjustments in favour of *wanna*). In both models, the greatest positive adjustment in favour of *wanna* is given to the contexts in which the subject is omitted, followed by the contexts with the personal pronouns *I* and *you* as the subject. Consider example (7) of a sentence with an omitted subject:

(7) Just wanna wrap this up now erm by bringing in the erm example of Greece. (DCJ, a lecture at a college).

Usually, it is *I* or *you* which are implied.

Finally, the second model based on the large dataset with imputed values contains an additional interaction, which is shown in Fig. 5. The effect of informativity of WANT given left context depends on informativity of WANT given the verb. It is negative when WANT is expected given the verb on the right, and slightly positive when WANT is not expected given the verb. That is, if a verb is faithful to this construction, the effect of the left context is strong and in the predicted direction. But if it is 'promiscuous' and occurs in many other contexts (e.g., *be* and *say*), then the effect of the left context is positive. This can be regarded as an anti-efficiency effect. Yet, the very broad confidence band associated with the promiscuous verbs tells us that we need more data to make a final judgement.

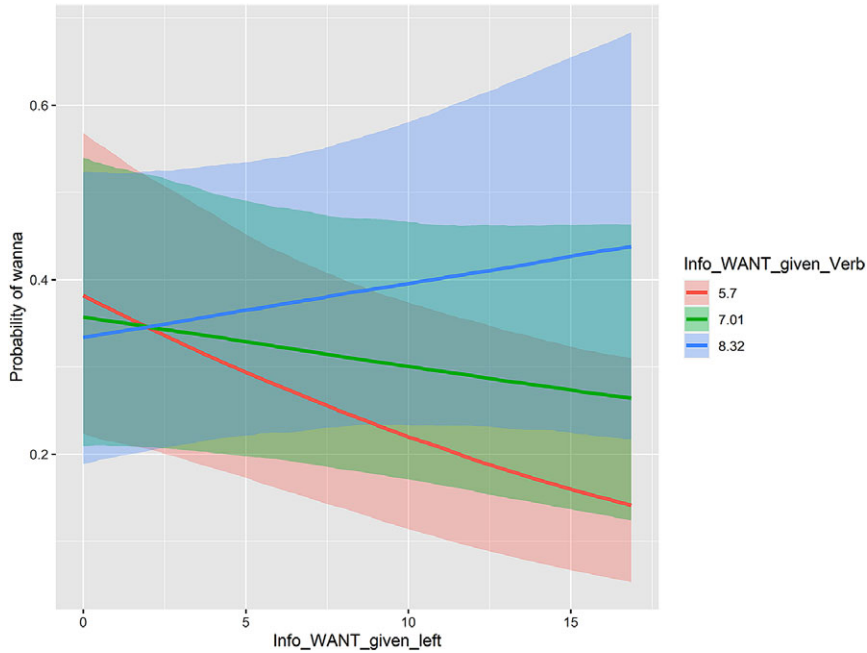

Fig. 5. Interaction between *Info_WANT_given_left* and *Info_WANT_given_Verb* in the model with imputed values. This is a conditional plot showing the predicted probabilities of *wanna* depending on Informativity of WANT given the word on the left and Informativity of WANT given the verb, computed for the following values of the predictors: conversations, female speakers, Age Group 0, no negation, not a question and mean values of the remaining numeric predictors in the dataset with imputed values. The lines correspond to the predicted posterior mean values, and the shaded areas are 95% error bands based on MCMC sampling. The blue line corresponds to the model predictions for high values of *Info_WANT_given_Verb* (8.32); the green line corresponds to medium values (7.01), and the red line corresponds to low values (5.7).

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

5. Conclusions

The main theoretical question we addressed was how the Principle of No Synonymy, which posits that all distinct forms also have distinct functions, can be reconciled with the bias towards efficient language use, by which the reduced and less costly form is chosen in more predictable contexts than the non-reduced and more costly one, and which assumes a certain degree of exchangeability of the forms. We presented two Bayesian mixed-effects models in order to test which contextual factors help us to predict the use of *want to* and *wanna* in the spoken component of the BNC. The factors included stylistic and sociolinguistic variables and different predictability measures, which served to test if the speakers used the variants efficiently. The first model was based on the small dataset without missing values, whereas the second model was based on the complete dataset with imputed missing values.

Before moving to the theoretical discussion, we should summarise our methodological finding. The first conclusion we can draw is that the results of the modelling based on complete observations and on the data with imputed values are very similar. The effect sizes and credibility levels of the predictors are nearly the same. One difference is that the use of the larger sample with imputed values allows us to discover another interaction between informativity of WANT given the word on the left and informativity of WANT given the infinitive. In the smaller sample, it did not play an important role. In addition, the large-sample model showed a clearer effect of the coordinate of a verb on Dimension 1 in Correspondence Analysis, which represents informality. Apparently, the use of the larger sample with a greater number of diverse verbs makes it possible to detect these effects. Finally, the larger model had narrower credibility intervals. This is not surprising: the more data a Bayesian model has (other things being equal), the more certain it is about the estimates.

In both models, most of the effects can be interpreted in terms of informality, which boosts the chances of *wanna.* In particular, *wanna* is more often preferred in conversations and with verbs that are commonly used in informal text types. In addition, the random effects suggest that *wanna* is most frequently used when the subject is a speech act participant (*I* or *you*), or is missing altogether, which serves as another indication of informality.

Wanna is also more frequently found in fast speech. This finding can be interpreted in two ways. On the one hand, high speech rate creates pressure for reduction. On the other hand, high speech rate is a property of causal speech, which means that it can capture some additional stylistic variation.

We also find expected patterns of sociolinguistic variation. If the speaker is young, he or she is more likely to use *wanna.* These results largely match previous findings from American English (Lorenz, 2013, 2020), suggesting that the emancipation of *wanna* from *want to* is an on-going process. Male speakers also use *wanna* more than female speakers do. This supports the observation that men use non-standard variants more often than women, who tend to prefer variants with high overt prestige (Labov, 1990; Romaine, 2003). The present data are coarse-grained in this respect, as the BNC is designed to represent British English as a whole. In smaller-scale speech communities the WANT variable might have more specific indexical values, which this study cannot capture.

Moreover, the full variant *want to* is preferred if there is a negative particle before WANT. Negation can be interpreted as a sign of greater cognitive complexity

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

(cf. Rohdenburg, 1996). Alternatively, negation is less frequent and therefore less expected than affirmation (Diessel, 2019, p. 228). This is why it can be more efficient to use the longer form *want to* in negative contexts. More research is needed in order to understand the origin of this effect.

We see thus that style and sociolinguistic variables play a central role, in accordance with previous research. *Wanna* has developed strong stylistic connotations, which now determine its usage. It is also strongly entrenched as a construction of its own with distinct properties, which nearly excludes its use as a pronunciation variant of *want to* in the same communicative context. Recall the title of the song discussed in Section 1, *Girls just want to have fun.* The variant *want to* appears in the official (written) title of the song. In the song itself, Cyndi Lauper pronounces *wanna.* So, the variants are interchangeable semantically, but not stylistically. They have strong associations with two different modalities.

This kind of divergence is most likely the reason why predictability has only a restricted effect on the use of the variants. *Wanna* is chosen more often when WANT is more likely to follow after the preceding word, in accordance with the expectations. However, this effect is observed only when the speech rate is high. This means that *wanna* helps to save articulation effort in fast speech when WANT is more predictable. Moreover, this effect is observed only for verbs that are more likely to be used after WANT, as the other interaction term suggests. Taken together, these conditions may represent a persistence effect of the origin of *wanna* in phonetic reduction. Reductions that level morphological boundaries often require favouring conditions in communicative and articulatory terms (such as predictability and rapid speech; cf. Lorenz & Tizón-Couto, 2020). It seems that the choice of *wanna* is still boosted when these factors conspire.

The mechanism explaining the relationships between predictability effects and the Principle of No Synonymy could be as follows. First, reduced forms of a construction arise due to the pressure for efficient communication in contexts where the construction is highly predictable. Next, two developments are possible. If the variants are not perceived as formal alternatives, the Principle of No Synonymy does not come into operation. The variants are then used interchangeably depending on the predictability. But if the variants are perceived as alternatives, the Principle of No Synonymy will pull them apart functionally.

Whether or not two variants are perceived as alternatives seems to depend on how salient the formal differences between them are. In very fast speech, *want to* and *wanna* (or any intermediate form) can be difficult to distinguish, which is why reduction can be employed for efficiency reasons. Due to individual differences, some uses will emerge in slower speech, as well. When a reduced form solidifies into a recognisably distinct variant and attains a certain frequency and salience, the variants are perceived as alternatives. Since *wanna* is a reduced form and used most often in fast speech, which is associated with informality, a natural path for differentiation of the variants is along the distinctions in the indexical field of stylistic and social stereotypes. This development is also strengthened by the secondary modal verb schema, which subsumes *gonna, gotta* and similar informal reduced forms (Krug, 2000; Diessel, 2019: Ch. 4).

More generally, the results support our hypothesis that predictability effects are less likely to be found if the variants have salient functional distinctions. When the variants are less stylistically and semantically contrastive, predictability effects can be easier to detect. For example, *that*-omission, which is not associated with salient

https://doi.org/10.1017/langcog.2022.7 Published online by Cambridge University Press

stylistic or semantic differences, strongly depends on predictability of the complement clause (see Section 1). The WANT alternation represents the opposite pole: the variants are highly distinct, and display hardly any predictability effects. It would be interesting to see where other alternations are located on this continuum, and whether there is a trade-off between functional distinctiveness and the role of predictability.

Supplementary Materials. To view supplementary materials for this article, please visit <https://doi.org/10.1017/langcog.2022.7>.

Acknowledgments. The first co-author's research presented in this paper was funded by the Netherlands Organisation for Scientific Research (NWO) under Gravitation grant 'Language in Interaction', Grant No. 024.001.006. We thank the reviewers for their detailed and constructive feedback, which has helped us improve the paper substantially. All remaining errors are ours.

Sources. *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Data availability statement. The datasets and R code are available in the following OSF online repository: https://osf.io/rd3jf/?view_only=1ff821da50d7420289b64fc275680797.

References

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schliperoord, & W. Spooren (eds), *Text representation*, 29–87. Amsterdam: John Benjamins.

Barth, D. (2019). Effects of average and specific context probability on reduction of function words BE and HAVE. *Linguistics Vanguard* 5(1), 20180055. https://doi.org/10.1515/lingvan-2018-0055

Bell, A., Brenier, J.M., Gregory, M., Girand, C. & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60, 92–111.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2), 1001–1024.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Boas, H. C. (2004). You wanna consider a constructional approach towards *wanna*-contraction? In M. Achard & S. Kemmer (eds), *Language, culture, and mind*, 479–491. Stanford: CSLI Publications.

Bolinger, D. (1977). *Meaning and form*. New York: Longman.

Bolinger, D. (1981). Consonance, dissonance and grammaticality: The case of *wanna*. *Language and Communication* 1, 189–206.

Bořil, T. & Skarnitzl, R. (2016). Tools rPraat and mPraat. In P. Sojka, A. Horák, I. Kopeček & K. Pala (eds), *Text, speech, and dialogue*, 367–374. Cham: Springer International Publishing.

Bresnan, J., Cueni, A., Nikitina, T. & Baayen, H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer & J. Zwarts (eds), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Broadbent, J. M. & Sifaki, E. (2013). *To*-contract or not *to*-contract? That is the question. *English Language and Linguistics* 17(3), 513–535.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, J., File-Muriel, R. J. & Napoleão de Souza, R. (2016). Special reduction: A usage-based approach. *Language and Cognition* 8, 421–446.

Bybee, J. & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4), 575–596.

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American Speech* 82 (1), 32–64.
Chambers, J. K. & Trudgill, P. (1998). *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.
Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (ed), *Mechanisms of language acquisition*, 1–33. Hillsdale, NJ: Lawrence Erlbaum Associates.
Cohen Priva, U. (2008). Using information content to predict phone deletion. In N. Abner & J. Bishop (eds), *Proceedings of the 27th west coast conference on formal linguistics*, 90–98. Somerville, MA: Cascadilla Proceedings Project.
Coleman, J. (2019). *Transcription textGrids for the audio edition of the British National Corpus 1993*. [Data Collection]. Colchester, Essex: UK Data Archive. https://doi.org/10.5255/UKDA-SN-851496
Coleman, J., Baghai-Ravary, L., Pybus, J. & Grau, S. (2012). *Audio BNC: The audio edition of the Spoken British National Corpus.* Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC.
Daugs, R. (2021). Contractions, constructions and constructional change: Investigating the constructionhood of English modal contractions from a diachronic perspective. In M. Hilpert, B. Cappelle & I. Depraetere (eds), *Modality and diachronic construction grammar*, 12–52. Amsterdam: John Benjamins.
Diessel, H. (2019). *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.
Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics* 12(4), 453–476.
Ellis, N. C. (2002). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition* 24(2), 297–339.
Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua* 142, 27–41.
Falk, Y. N. (2007). Do we *wanna* (or *hafta*) have empty categories? In M. Butt & T. H. King (eds), *Proceedings of the LFG07 conference*, 184–197. Stanford: CSLI Publications.
Ferreira, V. S. & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40, 296–340.
Finegan, E. & Biber, D. (2001). Register variation and social dialect variation: The register Axiom. In P. Eckert & J. R. Rickford (eds), *Style and sociolinguistic variation*, 235–267. Cambridge: Cambridge University Press.
Flach, S. (2020). Reduction hypothesis revisited: Frequency or association? In C. Sanchez-Stockhammer, F. Günther & H.-J. Schmid (eds), *Language in mind and brain: Multimedial proceedings of the workshop held at Ludwig–Maximilian University Munich, December 10–11, 2018*, 16–22. München: Open Access LMU.
Fowler, C. A. & Housum, J. (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26, 489–504.
Frank, A. & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae & V. M. Sloutsky (eds), *Proceedings of the 30th annual meeting of the cognitive science society (CogSci08)*, 939–944. Washington, D.C.: Cognitive Science Society.
Gardner, M. H., Uffing, E., Van Vaeck, N. & Szmrecsanyi, B. (2021). Variation isn't that hard: Morphosyntactic choice does not predict production difficulty. *PLoS ONE* 16(6), e0252602. https://doi.org/10.1371/journal.pone.0252602.
Getz, H. R. (2019). Acquiring *wanna*: Beyond universal grammar. *Language Acquisition* 26(2), 119–143.
Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L. & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Science* 23(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: Chicago University Press.
Goldberg, A. E. (2019). *Explain me this: creativity, competition, and the partial productivity of constructions*. Princeton: Princeton University Press.
Greenacre, M. (2007). *Correspondence analysis in practice*. London: Chapman & Hall.
Gries, S. Th. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement.* New York: Continuum.
Haiman J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language* 56(3), 515.
Hall, K. C., Hume, E., Jaeger, T. F. & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4(s2), 20170027. https://doi.org/10.1515/lingvan-2017-0027

Harrell, F. E. Jr. (2015). *Regression modeling strategies*. 2nd ed. Cham: Springer.

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics* 57(3), 605–633. https://doi.org/10.1017/S0022226720000535

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hollmann, W. & Siewierska, A. (2011). The status of frequency, schemas, and identity in Cognitive Sociolinguistics: A case study on definite article reduction. *Cognitive Linguistics* 22(1), 25–54.

Huddleston, R. & Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hudson, R. (2006). *Wanna* revisited. *Language* 82(3), 604–627.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology* 61(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology* 4, 230.

Jaeger, T. F. & Buz, E. (2017). Signal reduction and linguistic encoding. In E. M. Fernández & H. S. Cairns (eds), *Handbook of psycholinguistics*, 38–81. Hoboken, NJ: Wiley-Blackwell.

Jaeger, T. F. & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(3), 323–335.

Jurafsky, D., Bell, A., Gregory, M. & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (eds), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: John Benjamins.

Kaatari, H. (2016). Variation across two dimensions: Testing the complexity principle and the uniform information density principle on adjectival data. *English Language and Linguistics* 20(3), 533–558.

Krug, M. G. (2000). *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: Mouton de Gruyter.

Krug, M. G. (2001). Frequency, iconicity, categorization: Evidence from emerging modals. In J. Bybee & P. Hopper (eds), *Frequency and the emergence of linguistic structure*, 309–335. Amsterdam: John Benjamins.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Oxford: Elsevier.

Kurumada, C. & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83, 152–178.

Labov, W. (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2(2), 205–254.

Lakoff, G. (1970). Global rules. *Language* 46(3), 627–639.

Laporte, S., Larsson, T. & Goulart, L. (2021). Testing the Principle of No Synonymy across levels of abstraction: A constructional account of subject extraposition. *Constructions and Frames* 13(2), 230–262.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

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

Levshina, N. (2018). *Towards a theory of communicative efficiency in human languages*. Habilitation thesis, Leipzig University. https://doi.org/10.5281/zenodo.1542857.

Levshina, N. (in press). Comparing Bayesian and frequentist models of language variation: The case of help + (to) infinitive. In O. Schützler & J. Schlüter (eds), *Data and methods in Corpus linguistics – Comparative approaches*. Cambridge: Cambridge University Press.

Levshina, N. & Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard* 7(s3), 20200081. https://doi.org/10.1515/lingvan-2020-0081

Levy, R. & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt & T. Hoffman (eds), *Advances in neural information processing systems 19: Proceedings of the 2006 conference*, 849–856. Cambridge, MA: MIT Press.

Lorenz, D. (2013). *Contractions of English semi-modals: The emancipating effect of frequency*. Freiburg: NIHIN Studies/Universitätsbibliothek Freiburg.

Lorenz, D. (2020). Converging variations and the emergence of horizontal links: *To*-contraction in American English. In L. Sommerer & E. Smirnova (eds), *Nodes and networks in diachronic construction grammar*, 243–274. Amsterdam: John Benjamins.

Lorenz, D. & Tizón-Couto, D. (2020). Not just frequency, not just modality: Production and perception of English semi-modals. In P. Hohaus & R. Schulze (eds), *Re-assessing modalising expressions. Categories, co-text, and context*, 79–107. Amsterdam: John Benjamins.

Mahowald, K., Fedorenko, E., Piantadosi, S. T. & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126, 313–318.

Mair, C. (2017). From priming and processing to frequency effects and grammaticalization? Contracted semi-modals in Present-Day English. In M. Hundt, S. Mollin & S. E. Pfenninger (eds), *The changing English language: Psycholinguistic perspectives*, 191–212. Cambridge: Cambridge University Press.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

Nesselhauf, N. (2014). From contraction to construction? The recent life of *'ll*. In M. Hundt (ed), *Late modern English syntax*, 77–89. Cambridge: Cambridge University Press.

Nicenboim, B. & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas - Part II. *Language and Linguistics Compass* 10, 591–613. https://doi.org/10.1111/lnc3.12207

Okazaki, M. (2002). Contraction and grammaticalization. *Tsukuba English Studies* 21, 19–60.

Piantadosi, S. T., Tily, H. & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9), 3526–3529.

Postal, P. M. & Pullum, G. K. (1982). The contraction debate. *Linguistic Inquiry* 13(1), 122–138.

Pullum, G. K. (1997). The morpholexical nature of English *to*-contraction. *Language* 73, 79–102.

R Core Team (2020). R: A language and environment for statistical computing. Version 4.0.2. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Raymond, W. D., Dautricourt, R. & Hume, E. (2006). Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18(1), 55–97.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2), 149–182.

Romaine, S. (2003). Variation in language and gender. In J. Holmes & M. Meyerhoff (eds), *The handbook of language and gender*, 98–118. Maiden, MA: Blackwell.

Sag, I. A. & Fodor, J. D. (1994). Extraction without traces. *West Coast Conference on Formal Linguistics* 13, 365–384.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells. From corpus to cognition*. Berlin, New York: Mouton de Gruyter.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1), 140–155.

Staum, L. (2005). When stylistic and social effects fail to converge: A variation study of complementizer choice. Ms. Available at http://staum.casasanto.com/documents/StaumQP2.pdf.

Szmrecsanyi, B. & Hinrichs, L. (2008). Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In T. Nevalainen, I. Taavitsainen, P. Pahta & M. Korhonen (eds), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: John Benjamins.

Tagliamonte, S. A. & Roeder, R. V. 2009. Variation in the English definite article: Socio- historical linguistics in t'speech community. *Journal of Sociolinguistics* 13(4), 435–471.

Tagliamonte, S. A., Smith, J. & Lawrence, H. (2005). No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change* 17(1), 75–112.

Tizón-Couto, D. & Lorenz, D. (2018). Realizations and variants of *have to*: What corpora can tell us about usage-based experience. *Corpora* 13(3), 371–392.

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Uhrig, P. 2015. Why the Principle of No Synonymy is overrated. *Zeitschrift für Anglistik und Amerikanistik* 63 (3), 323–337. https://doi.org/10.1515/zaa-2015-0030

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3), 1–67. https://www.jstatsoft.org/v45/i03/.

Vasishth, S., Chen, Z., Li, Q. & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE* 8(10), 1–14. https://doi.org/10.1371/journal.pone.0077006

Wasow, T., Jaeger, T. F. & Orr, D. M. (2011). Lexical variation in relativizer frequency. In H. J. Simon & H. Wiese (eds), *Expecting the unexpected: Exceptions in grammar*, 175–195. Berlin: De Gruyter Mouton.

Zipf, G. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley Press.

Cite this article: Levshina, N. & Lorenz, D. (2022). Communicative efficiency and the Principle of No Synonymy: predictability effects and the variation of *want to* and *wanna Language and Cognition* 14: 249–274. https://doi.org/10.1017/langcog.2022.7